ELEMENTS OF STATISTICS

ELEMENTS OF

ST·ATISTICS

ARTHUR L. BOWLEY, C.B.E., Sc.D.

Emeritus Professor of Statistics in the University of London

SIXTH EDITION

S T A P L E S   S   P R E S S

PREFACE TO FOURTH EDITION

This book was first published in 1901, and was then based on lectures delivered at the School of Economics in the five years following its foundation in 1895. Two further editions have been issued in which the text was revised without any important alteration, and an Appendix added dealing with the second approximation to the normal curve of error, and subsequently some pages of addenda were circulated. In the present edition Part I remains substantially as it was in 1901, except that Section III of Chapter III has been replaced by a new illustration, the chapter on Averages has been rearranged, a chapter on the measurement of dispersion takes the place of the former Chapter V, in Chapter IX the treatment of retail index-numbers has been reconsidered, and the second section of Chapter X has been recast. At the same time those parts of the text which were out of date have been replaced by more modern material and the whole has been revised, but with as little alteration of the original as possible, since a revised version may by too much attention to detail destroy the balance of the original. On the other hand, Part II has been completely rewritten and considerably extended, both by the more detailed and extended treatment of theory and by the addition of a number of examples which illustrate the arithmetical use of the formulæ and show the scope of the application of the theory. For the convenience of those who possess the earlier edition, to whom the revised Part I contains little that is new, Part II is issued separately; while for those whose mathematical knowledge is too slight to allow them to follow the treatment in Part II in its new form Part I is also issued separately.* But the two Parts together are essentially one book with a common index and with cross references from one to another.

The whole book is intended to form a general introduction to the theory and practice of statistics for all persons whose business it is to handle them or to whom a general understanding both of the utility of statistical results and the limitations

* The Parts are no longer issued separately.—April, 1926.

of statistical investigation is important. It is not in any way intended to be a compendium of facts, and the tables inserted are only to afford illustrations of method, nor does it contain any detailed account of published statistics; but it is hoped that a reader will find himself in a position to understand, and above all to appraise and criticise, tables and results published officially or otherwise relating to any of those very numerous subjects in which numerical knowledge of facts and their inter-relation is essential. No attempt is made to treat the history or bibliography of the subject; there are many books extant in English, French and German which devote considerable space to the historical development of the methods and practice of statistics, with bibliographical references; it seemed better here to omit these aspects altogether than to give them a cursory treatment. With these limitations it is hoped that the treatment in Part I covers adequately the great part of the methods and technique necessary for ordinary statistical work so far as this can be done without the use of any but the most elementary mathematics. The chapter on Interpolation, indeed, uses symbols which at first sight may look formidable to the non-mathematician; but in fact the use of finite differences and of Newton's formula of interpolation is quite simple and the arithmetic involved very easy, and the great part of the chapter should be readily intelligible to those who have a school training in graphic algebra.

Part II makes much greater demands both on preliminary training and on the power of following somewhat involved abstract reasoning. The actual knowledge postulated is that obtainable in a graduate course on the calculus, and the only theorems not generally included in such a course are proved (in an abbreviated form) in the Appendix. In the first edition an effort was made to obtain the principal results without the use of the Calculus; but as the subject has developed during the past twenty years, it has become necessary to abandon this attempt. The results that can be reached by algebra alone are no doubt important and useful, but there is so much of at least equal utility that can only be appreciated after more advanced mathematical study that a student will save time in the end by becoming familiar with the elements of the infinitesimal calculus before he commences the serious study of mathematical statistics. This opinion is confirmed by the very

loose reasoning often employed by writers who make too facile use of the standard deviation, of curves of frequency and especially of the coefficient of correlation. Very great care has been taken in Chapter VI, Part II, to show as exactly as possible the meaning of the measurement of correlation by this coefficient and its implications, and very much more might have been said before the subject was too thoroughly explored. No one should attempt to measure correlation till he has studied the theory closely and critically.

Though the treatment in Part II is intended to serve as a general introduction to mathematical statistics whatever the subject-matter to which they are applied and to include definitions and explanations of the terms and measurements in common use, so as to be of assistance to students in all branches of science that involve group measurements, yet the order of treatment and in particular the worked examples are chosen principally with reference to the problems that arise in sociological and economic investigations, many of the examples in fact being taken from researches I have personally made in which mathematical treatment was only introduced so far as the line of inquiry called for it. In consequence of this the reader who is familiar with the writings of Professor Karl Pearson, Mr. Elderton, Mr. Hardy, Mr. Yule and Dr. Greenwood will notice that little emphasis is laid on applications to biological or to actuarial problems, while prominence is given to formulæ and to methods which have received less attention.

It is unfortunately the case that a great deal of controversy has arisen with reference not only to the best methods of treatment, but also to the fundamental conceptions that underlie the application of the principles of mathematical probability to statistical observations. I cannot hope to have avoided controversial questions (for, indeed, if these were rigidly excluded there would be little left), but I have endeavoured to put in the foreground those methods and principles which command general acceptance and to omit those which are the subject of dispute and are unessential. In one respect, however, a definite course is followed which will not meet with universal approval; in my opinion the standard deviation has only limited utility unless it is connected with a table of probability by which the chances of exceeding given multiples of this deviation can be calculated, and consequently I have emphasised the normality

Formulæ for the standard deviation of the mean difference for frequency groups were worked out by me, with Mr. R. G. D. Allen's help, in February 1936, and communicated to the International Congress of Mathematicians at Oslo in July. During the Congress Mr. H. Wold found a more direct way of obtaining them. Only that relating to the normal curve is here given (p. 487).

It is feared that the Supplement will not provide easy reading. It is deliberately compressed without omitting essentials. Here I can only say that an attempt has been made to simplify the treatment of problems in the articles or books in which they were first considered, and to avoid the use of mathematical methods which are unfamiliar to the non-expert.

A. L. B.

Marley Hill,
January 1937.

CONTENTS

PART I
GENERAL ELEMENTARY METHODS

xi

PART II

APPLICATIONS OF MATHEMATICS TO STATISTICS

SUPPLEMENTS

CONTENTS

', LIST OF DIAGRAMS ,

PART I

PART II

SUPPLEMENT. ,

PART I.

GENERAL ELEMENTARY METHODS.

.

PART I.

GENERAL ELEMENTARY METHODS.

CHAPTER I.

SCOPE AND MEANING OF STATISTICS.

VERY many definitions have been given of the word *statistics*, and each author who has written on the subject has assigned new limits to the field which should be Definitions of statistics. included in its scope. It will not be necessary for the purpose of this book to discuss the merely verbal differences involved, but only to explain what is intended by its title, and to consider the limits of the science which it is proposed to investigate. It will be useful, however, to mention some possible definitions.

Statistics may, for instance, be called *the science of counting*. Counting appears at first sight to be a very simple operation, which any one can perform or which can be done The science of counting. automatically; but, as a matter of fact, when we come to large numbers, *e.g.*, the population of the United Kingdom, counting is by no means easy, or within the power of an individual; limits of time and place alone prevent it being so carried out, and in no way can absolute accuracy be obtained when the numbers surpass certain limits. Great numbers are not counted correctly to a unit, they are estimated; and we might perhaps point to this as a division between Distinction between statistics and arithmetic. arithmetic and statistics, that whereas arithmetic attains exactness, statistics deals with estimates, sometimes very accurate, and often sufficiently so for their purpose, but never mathematically exact. Statistics generally relate to numbers so great that their estimation is beyond the

power of an individual, and requires the co-operation of
Statistics as co-operative counting. an organised body of workers. Though the
collection of numbers by several persons and
the mere addition of the results seem simply
questions of arithmetic, yet in practice two difficulties soon
occur.   First, it is not easy to define the thing to be counted
so explicitly that all the tellers shall admit and reject instances
on the same principles; for such simple objects as the number
of rooms or stories of a house, a person's age, even an indi-
vidual, give rise to such complex questions of definition that
it is often impossible to tell from a short description of a
category exactly what items are included in it.   Secondly,
numerical errors cannot be avoided when many workers are
involved; for some among a large number of persons will be
inaccurate, some unintelligent, some will not obtain complete
information, and when their reports are compiled there will be
occasional mistakes in copying and errors in tabulation.   A
total which is the result of the work of many hands will cer-
tainly from one cause or another fall short of complete accuracy.
But though all estimates of this nature are sometimes included
under the term *statistics*, this definition at once is too wide,
and also does not bring out the distinctive nature of statistical
method.

It is better, in fact, to define statistics *a posteriori*.   In
dealing with masses of figures, large numbers descriptive of
Statistics as a method. groups, series of totals or averages relating to
different dates or places, it is found that special
methods become necessary—methods which depend on par-
ticular properties of large numbers, methods which are suitable
for describing complex groups so that they can be easily com-
prehended, methods for analysing the accuracy of statements,
for measuring the significance of differences, for comparing one
estimate with another.   Those estimates to which these
methods apply are within the scope of statistics; it is the
study of these methods that is the object of this book.   It is
clear that, under our tentative definition, statistics is not
Generality of statistical method. merely a branch of political economy, nor is it
confined to any one science.   A knowledge of
statistics is like a knowledge of foreign languages
or of algebra : it may prove of use at any time under any
circumstances.

It may be interesting to trace the connection of statistical method with various branches of knowledge. To begin with the physical sciences : there are two points in Its use in the which this method touches astronomy. The physical sciences. method of least squares was introduced by an astronomer, anxious to choose the best of several slightly discrepant observations of the position of a star. In most physical observations several measurements are taken of the same quantity, and it is found that, however carefully they are made, they never absolutely agree; just as the averages obtained by different statisticians from the same series of sociological observations are generally not identical. From such a group of measurements it is necessary to deduce the most probable estimates; this is done by the application of the law of error, in the form of the method of least squares.

The other point of resemblance of statistical to astronomical method is common also to geology and to most applied sciences. The course of scientific measurement Progressive has generally been to take first a rough observation accuracy. of a quantity, such as the distance of the sun, the thickness of a stratum, the atomic weight of an element, the specific gravity of a substance; then, as information accumulated, as the precision of instruments increased and methods were better adapted, to make the measurement gradually more and more accurate. It is important to appreciate this development, for in the present state of our knowledge, many statistical measurements cannot be made with precision for want of data, and a critic is inclined to say that for this reason preliminary estimates are valueless; but from the scientific point of view this criticism is wrong, for a faulty measurement made on logical principles is better than none, if limits can be assigned to its possible error, and may lead to others with progressive improvement.

Passing by the general resemblance of statistical investigations to all scientific experiments, we may notice the use of statistics in biology. It was, perhaps, not recog- Statistics and nised before the publication of Professor Karl biology. Pearson's investigations,* that the whole doctrine of evolution and heredity rests in reality on a statistical basis. It is in

---

* See *The Grammar of Science*, 1900, chap. x. *seq.*, and the references there given.

this direction that some of the most important new work in mathematical statistics is being done. It may be worth while to sketch very briefly the nature of the problem. Out of a great number of observations, say the measurements of the heights of a group of men, the type is found—an average, about which all the measurements are grouped according to some definite law. The problem is then to determine whether this type or the grouping about it changes, and in what way. The differences found in successive generations form the data on which arguments as to evolution and development are founded. The method applies equally to fossil remains, to zoological species, and to many other groups. If it is neglected, many valid arguments lose a great part of their force, and theories are founded on personal impressions of phenomena instead of on scientific measurement. The work done in this direction becomes of immediate use to the student of social questions. The average wage and the grouping about it and the change in these quantities present precisely similar problems; the correlation between the effects of different factors are calculated by the same mathematical formulæ; in fact, these methods furnish the only accurate way of measuring numerical changes in complex groups. Much valuable information has been collected in anthropometrical laboratories, which has increased the statistician's knowledge of facts and given birth to important theoretical principles.

Meteorology has much in common with statistics. The chief measurements taken for the purposes of this science are Statistics and of temperature, barometrical pressure, moisture meteorology. of the air, and force of the wind. One of the problems attacked is again that of finding the type from a group of observations, and of measuring its change. The tables which state the average temperature year by year are in many ways similar to those which the Registrar-General publishes of births, deaths, and marriages. Without the aid of statistical method, the averages obtained show mere numbers from which no logical deductions can be made. With the help of this knowledge, it can be seen whether the change from year to year is significant or accidental; whether the figures show a progressive or periodic change; whether they obey any law or not. The problem is easily seen to be of importance for forecasting the future population and for many similar purposes.

We are thus brought by a short step to the province to which statistics has sometimes been confined : the study of demography. If in demography we include, not merely the measurement of the numbers of the population, the birth, marriage, and death rates, the distribution by age, by sex, and by locality, in fact, the figures which naturally come from the census and the Registrar-General's returns; but include also, industrial and social measurements, of distribution of the population by trade, of income, wages, prices, production, foreign trade, transport, and so forth ; we have extended the limits of demography till it includes the majority of the statistical investigations directly interesting to students of sociology or of political economy. Without stopping to decide the exact limits of demography, we can quickly pass to another definition of statistics (so far as it concerns such students) on which it is wished to lay a certain stress : *statistics is the science of the measurement of the social organism, regarded as a whole, in all its manifestations.* In a monograph, after the fashion of Le Play, a single family is studied; the occupations and earnings of its members, the way these earnings are spent, and its economic position generally are set down ; but this study is not so far statistical. In demography we study the same quantities when groups of families are concerned; the number of families engaged in certain industries, and their average receipts, expenditure, and savings ; here we have statistics. In the monographic method the individual is everything ; in the statistical method, nothing. When we wish to obtain a measurement of the group, peculiarities of individuals receive no attention; it is only when the same peculiarities are possessed by many persons that they become of importance. Statistics may rightly be called the science of averages. In the measurement of a complex group, say of incomes and wages, the exceptional artiste who can earn £100 in an evening, and the inefficient labourer who can only make sixpence a day, affect only slightly the general average; they are not entered in separate categories ; but the large group of skilled artisans who earned before 1914 forty shillings a week, or of casual labourers who made less than fifteen shillings, are entitled to separate notice. The exact specification to be adopted is only a question of degree, which differs with the

Statistics and demography.

Statistics relate to the social organism as a whole.

nature of the particular investigation in hand. The object of a statistical estimate of a complex group is to present an outline, to enable the mind to comprehend with a single effort the significance of the whole. To do this it is necessary to exclude rigorously any presentation of details, for the same reason that, in a painter's rendering of a tree, the individual leaves are not distinguished. The outline will be a little blurred, a little inaccurate; but it will be as distinct and detailed as the mind has power to grasp it, or the eye to see it; the impression will be rightly given. There is a very important principle involved in this method. The individual members of a group vary continually, the whole group varies very slowly. It is impossible to follow or measure the motions of separate atoms; it is comparatively easy to state the laws of motion for a solid body. Great numbers and the averages resulting from them, such as we always obtain in measuring social phenomena, have great inertia. The total population, the total income, the birth and death rates, average wages, change very little; similar quantities relating to a single family change very fast. It is this constancy of great numbers that makes statistical measurement possible. It is to great numbers that statistical measurement chiefly applies.

The relation of statistics to political economy is a simple one. Professor Marshall says,* " Statistics are the straw out

Statistics and political economy. of which I, like every other economist, have to make the bricks." The statistician furnishes the political economist with the facts, by which he tests his theories or on which he bases them. Since the economist deals chiefly with phenomena relating to groups, and regards the individual only as a member of a group, it is to statistics as the science of averages that he looks for his information. When he is dealing with national economy, with the volume of trade, for instance, or the purchasing power of money, he is limited to pure theory, till statistics as the science of great numbers has provided the facts. The chemist experimenting in his laboratory is like the statistician; the chemist theorising in his study is like the economist. Because of this relation it may be held to be the business of the statistician to collect, arrange, and describe, like a careful experimentist, but

---

' Evidence to the Committee on the Census, 1890.

to draw no deductions; even in an investigation relating to cause and effect, to present evidence but not conclusions.   As a distinct operation, of course, the statistician may assume the rôle of the economist, for the same man may well be qualified to conduct the experiment and fit the theory.   And just as a theoretical chemist will have little or no power unless he fully appreciates experimental methods and difficulties, even if he has not the manual dexterity to conduct them to perfection himself, so no student of political economy can pretend to complete equipment unless he is master of the methods of statistics, knows its difficulties, can see where accurate figures are possible, can criticise the statistical evidence, and has an almost instinctive perception of the reliance that he may place on the estimates given him.

The proper function, indeed, of statistics is to enlarge individual experience.   An individual is limited to what he can himself see, a very small part of one division of the social organism; his knowledge is extended in various ways, by the conversation of his acquaint- *Statistics versus individual experience.* ance, by newspaper reports, by the writings of experts.   According to his ability and power of judgment, he will be able to form a correct view of the numerical importance of groups of persons and things; but it is in the highest degree improbable that he will not have been biassed by the peculiarities of his position, and that he will place his different items of information in the right perspective; and he will not be able to gauge rightly the accuracy of his data.   As soon as he begins to examine these points he is undertaking a statistical investigation, and will very soon find himself involved in all the difficulties and problems from which a knowledge of statistical method alone can disentangle him.   This is the obvious answer to those who deny the use of statistics.   A statistical estimate may be good or bad, accurate or the reverse; but in almost all cases it is likely to be more accurate than a casual observer's impression, and in the nature of things can only be disproved by statistical methods.   •

A chief practical use of statistics is to show relative importance, the very thing which an individual is likely to misjudge. Statistics are almost always comparative.   The *Statistics are comparative.* absolute magnitude ·of a quantity is of little meaning to us till we have some similar quantity with which

to compare it.   A statement of the number of paupers in the
United Kingdom is valueless unless we know the total popu-
lation.   A statement of the number of gallons of water supplied
per head to the people of East London is of little meaning to
us till we know the quantity supplied to other towns.   The
average wage, shown in the Wage Census, does not convey its
full significance till we have similar computations for other
countries or relating to other years.   In the case of most
statistical estimates, it will be found that we need another for
comparison before we can appreciate the meaning of the first.

If the group of objects which we wish to measure is large,
its enumeration will be beyond our unassisted efforts, or those
Official of any organisation at our command.   Some
statistics: investigations, indeed, have been successfully
conducted by private organisations, for instance, those which
resulted in Booth's *Life and Labour of the People*, Leone Levi's
Wages and Earnings, and Rowntree's *Poverty ;* and the method
of samples has also been used (*e.g.*, in *Livelihood and Poverty*,
by the present author and Burnett-Hurst) to reduce an inquiry
to manageable dimensions ; but in general the measurement
of a part of the social body or industrial organism must be
undertaken by the central or local governments, if it is to be
successfully carried out.   The fact that this is the case explains
the heterogeneity and the imperfection of the mass of statistics
extant.   A government primarily collects numerical informa-
tion only in relation to its own functions.   Thus the administra-
tion must know the numbers of the population and the area of
the country in gross and in detail for its own purposes.   Large
groups of figures come simply from the necessity of public
account-keeping.   Many official figures are bye-products ;
for office purposes an account is kept of all transactions in
which the government has a hand, and of industries subject
to special regulations ; and the government publishes most
of the figures which thus come in its way.   To such causes
have been due our knowledge of the statistics of income,
education, imports, 'railways, mines, factories, and so' on.
Though few figures are collected simply for scientific purposes,
yet in many cases schedules issued for administrative ends are
used at the same time for the reception of other information,
of use chiefly to the sociological student ; much of the Census
information comes under this heading.   A view of those

figures, relating to the United Kingdom, which are easily accessible to the student, can be obtained by turning through the annual *Statistical Abstract for the United Kingdom*, the *Annual Abstract of Labour Statistics*, and the *Registrar-General's Annual Report;* in one or other of these, summaries of, and references to, most official statistics are to be found.

It is clear that figures collected simply in connection with administrative purposes are not likely to be precisely those which are needed by the student of sociology or   their political economy. Even where the wants of the incompleteness. official and the student are nearly identical, the classification and tabulation may not meet scientific requirements. There has, indeed, been considerable progress in recent years, in the direction of amassing statistical information not absolutely needed by the administration, and much of the work of the Labour Department of the Board of Trade (now merged in the Ministry of Labour) was of this kind; but very much more might reasonably be done, at an expense which would be almost negligible when considered in relation to the national income. Thus the census might be made, in part at least, quinquennial, and the body of workers, who are organised once in ten years to conduct it, only to be disbanded when the report is issued, might be made permanent and entrusted with the carrying out of other inquiries on a national scale. Market and retail prices of many staple commodities could be tabulated, analysed and published. Movements of goods by rail could be tabulated in the same way as transport by water, and the anomaly that we know more of our foreign than of our home trade be removed. Records of home production need not be confined to agriculture, mining, and steel works, but extended on the lines of the Census of Production of 1907 till we know every year the output of the principal industries. Above all a central statistical office is needed which should co-ordinate all existing statistics and, working directly or through the appropriate Departments, aim at completing and perfecting a continuous statistical account of the nation. It needs very little study of statistics or of political economy   Central to feel the pressing need of more and better statistical office. co-ordinated information; illustrations of the gaps in our knowledge are easily found. When dealing with our national income we can obtain statistics of wages, and of income subject

to tax; but for salaries below the exemption limit, and for part of the income received from foreign investments, we are forced to rely on educated guesses. For the change of the purchasing power of money we know, thanks chiefly to the *Economist* and trade newspapers, the course of wholesale prices, but many interesting calculations are brought to a standstill because of the imperfection of the records of retail prices. With regard to wages, we can estimate fairly accurately standard and average wages, but, in default of an industrial census, do not know how many persons are in receipt of each given wage, nor the relative numbers of masters and men. Till there is a public demand for such information, it will need a very enlightened government to spare the time, trouble, and the relatively small sums of money necessary for a systematic attempt to fill up these gaps; but every one can do something towards this enlightenment, and in furtherance of this demand, by studying what has been done in other countries, and building up a knowledge of the science of statistical investigation.

The absence of such a demand is perhaps due to a widely spread and not unreasonable distrust of statistical estimates, *Distrust of statistics:* crystallised in the common remark that " anything can be proved by statistics." This is to a great extent the fault of the criticising public themselves: they are always requiring and the newspapers always supplying information, which depends on a statistical basis, but for which good statistics are not to be found for one or other of *its causes.* the reasons already indicated. The informant must perforce turn to inaccurate estimates, and the public has no knowledge or discrimination as to what estimates rest on satisfactory data, or indeed as to what quantities are capable of statistical evaluation. Again, figures which cover only part of the subject, such as the Wage Census average, or the *Labour Gazette* returns of unemployed, may be quoted as universal; mere estimates, made for quite other purposes, may be given as accurate and complete; and on such unreliable premises arguments are based, which naturally, by a judicious choice of material, can be made to support any theory at pleasure. It will generally be found that the statistician, on whose authority such statements are supposed to be based, is not to blame. Some of the common ways of

producing a false statistical argument are to quote figures without their context, omitting the cautions as to their incompleteness, or to apply them to a group of phenomena quite different to that to which they in reality relate; to take estimates referring to only part of a group as complete; to enumerate the events favourable to an argument, omitting the other side; and to argue hastily from effect to cause, this last error being the one most often fathered on to statistics. For all these elementary mistakes in logic, statistics is held responsible.

Perhaps statisticians themselves have not always fully recognised the limitations of their work. At best they can measure only the numerical aspect of a pheno- *Limitations of* menon; while very often they must be content *statistics.* with measuring not the facts they wish, but some allied quantity. We wish to know, for instance, the extent of poverty, its increase or diminution : poverty we cannot define or measure, and we cannot even count the number of the poor ; all we can do is to state the number of officially recognised paupers, and add perhaps some estimates from private sources ; but this gives us no clue to the intensity of poverty in individual cases. Or we wish to obtain statistics of health : but the principal measurements made are of the death-rate and average length of life, and the prevalence of some diseases, very different matters. The statistician's contribution to a sociological problem is only one of objective measurement, and this is frequently among the less important of the data ; it is as necessary, however, to its solution as accurate measurements are for the construction of a building.

CHAPTER II.

THE GENERAL METHOD OF STATISTICAL INVESTIGATION.

AT first sight it will seem as if there were no method common to all statistical investigations, and indeed the processes differ so widely that it is not easy to outline a scheme which will include them all; but the following sequence is generally indicated * as of general application, and will serve at least to thread an examination of methods together : (1) the Collection of Material, (2) its Tabulation, (3) the Summary, and (4) a Critical Examination of its results. The first three processes will be discussed in detail in the following chapters.

It may be well to state what equipment is necessary for the student who wishes to learn statistical methods. In collection

Preliminary knowledge necessary or expedient.

and tabulation common-sense is the chief requisite, and experience the chief teacher; no more than expertness in quite simple arithmetic is necessary for the actual processes; but since, as we shall see immediately, all the parts of an investigation are interdependent, it is expedient to understand the whole before attempting to carry out a part. For summarising, it is well to have acquaintance with the various algebraic averages, and with enough geometry for the interpretation of simple curves; though all the operations can be performed without the use of algebraic symbols. For criticism of estimates and interpretation of results, it is necessary to use the formulæ of more advanced mathematics, and it is obviously expedient to understand the methods by which these formulæ are obtained to ensure their intelligent use. They are specially necessary for the comparison of complex groups, and for estimating the significance of a divergence from the average, or the deviations in a list of periodic figures, and quite essential in dealing with correlation.

* See, e.g., Dr. Bertillon's *Cours élémentaire de Statistique*, to which the present author is indebted for some of the treatment in the following pages

14

(1) Information is generally collected by issuing blank circulars, forms of inquiry, to be filled in either by a few officials or by many individuals, and the proper drawing up of this form is one of the chief tasks in a good investigation. Before this form is issued it is necessary to formulate a complete scheme of the whole undertaking, and even to have some idea of what the resulting figures will be, so as to be able to arrange the details of the organisation on the right scale, and adjust the tools used to their purpose. As already pointed out, the object whose measurement is wanted is not in general exactly that which can be measured, and the measurable quantity nearest to it must be found; *e.g.*, when the average annual earnings of the working class were in question, the quantity first measured was the average weekly wage. Then some technical knowledge of the particular subject is needed; and, if not possessed, a preliminary inquiry on a small scale may be necessary to show how to fit means to ends. The people who possess the information required must be discovered and interrogated at first hand. The questions put must be those which will yield answers in a form ready for tabulation, and the scheme of tabulation must therefore be thought out beforehand. The questions must be so clear that a misunderstanding is impossible, and so framed that the answers will be perfectly definite, such as a simple number, or " yes " or " no." They must be such as cannot give offence, or appear inquisitorial, or lead to partisan answers, or suppression of part of the facts. The mean must be found between asking more than will be readily answered and less than is wanted for the purpose in hand. The form must contain necessary instructions, making mistakes difficult, but must not be too complex. The exact degree of accuracy required, whether the answers are to be correct to shillings or pence, to months or days, must be decided. Every word and every square inch of space must be keenly criticised. A little trouble spent upon the form will save much inconvenience afterwards.

Collection : blank forms ;

nature of the questions.

(2) In considering what method is to be adopted for tabulation, we must remember that the investigation is intended to furnish the answers to certain definite questions— how many people, what wage, what price—and each column must present some total which is relevant to these

Tabulation.

questions. The exact scheme employed will differ in different inquiries. In the population census, much of the tabulation is almost automatic; in the wage census, the best and simplest. way to show the grouping about the average wage in each occupation had to be specially devised; in trade statistics the number of different categories to be adopted and the limits of each raise difficult questions. In general, the scheme of investigation requires knowledge of certain groups; and the totals resulting from tabulation should show the numbers of items in these, so that after tabulation, instead of the chaotic mass of infinitely varying items, we have a definite general outline of the whole group in question.

(3) When the raw material is worked up to this point, skill of a different kind is wanted. From the numbers obtained, we Averaging and have to pick out the significant figures; so to summarisation. present the totals and averages as to give a true impression to an inquirer; to summarise briefly the information obtained; to concentrate the mass into a few significant averages, and to describe their exact meaning in the fewest and clearest words, for it is the result of this con- centration which will generally be used and quoted. To do this skilfully requires an acquaintance with the method of averages and the use of diagrams. It may further be necessary to fill in unavoidable gaps in the figures in order to supply esti- mates for intermediate years; this needs a study of the danger- ous method of interpolation. Finally, a verbal description of the process, its genesis and results, and an estimate of its accuracy must be written, and then the investigation is complete.

(4) The student who has to make use of statistics should not be content to take the results of an inquiry on authority, but Criticism of ought to acquaint himself with all these details of results. method. Before the results can be criticised, it is necessary to know the complete genesis of the figures; whether the whole field was covered; exactly whence the information tabulated was obtained; whether there was a possibility of bias; how nearly the individual answers were correct; whether the informants really knew the facts they related, and if they were likely to state them correctly. The published statement of the results should show clearly the whole scheme of collection so as to make this criticism possible; in particular, specimens of the original blank forms should be

included, so that the reader can judge whether the original answers lead definitely and exactly to the tabulated results. Internal evidence often leads to much useful criticism. It can be seen whether the number of returns for each group is proportional to its importance, or if a specially important figure depends on only slight evidence. The continuity of the figures can be examined, and the causes of sudden gaps investigated. The returns can be divided into sample groups, and the extent of the correspondence of these groups with the general result will often indicate whether the returns are sufficiently general. A careful study of the more minute tabulations may show within what percentage the final numbers may be expected to be correct. A critical examination of this kind will often show that the information obtained is insufficient to lead to precise results, and then attention should be directed to estimating the magnitude of the effect of omissions and inadequacy of data.

A most important function of statistics is to produce evidence showing the relation of one group of phenomena to another; for the information obtained is presumably intended as a guide for action, the guidance is generally needed to show what actions are likely to produce certain desired effects, and this is best investigated by finding how such effects have been produced in the past. We have then to determine whether changes in one measurable quantity have produced changes in another; a problem very often insoluble, but one on which most light can be obtained by the study of the relevant statistics in the light of mathematics, the mathematics of probability, and it is in this particular branch of mathematics that recent statistical progress has been chiefly made.

Such questions, however important, are somewhat abstruse, and presuppose a certain amount of technical knowledge which is not in the possession of the general student. The plan of this book is to postpone all questions requiring such technical or mathematical knowledge to the Second Part, and to confine our earlier discussions to problems needing no special training or equipment.

CHAPTER III.

DEFINITION OF UNIT. COLLECTION OF DATA.

Preliminary. Definition of Unit.

ALMOST the first question in the initiation of an investigation is, What is to be counted?, and nearly the last question when the tabulation is completed is, What has been counted? The answer to the former gives the preliminary definition, that to the latter shows how it had to be modified in practice. The essential difficulties of definition come, first, from the need of interpreting conceptions conveyed or obscured by ordinary words into entities capable of enumeration, and, secondly, from determining the things that can actually be counted which are nearest to the entities of which knowledge is desired. Thus we may be investigating overcrowding or loss Quæsita and of work through unemployment. Overcrowding data. is expressed numerically in the relation between persons and room or air-space, and differs with the age and sex of the members of the household and the ventilation and light of the rooms. In practice, persons only can be counted (without detailed reference to their needs), and the number of rooms (a room being defined rather arbitrarily) or their cubic contents can be recorded. Loss of work is expressed numerically in the number of ordinary working days on which no paid work was done. In practice, those are counted as unemployed who satisfy certain formalities at trade union or Labour Exchange offices, such as signing a register at a particular hour each day. The definition of "number unemployed" depends on the regulations relating to these registers, and among unemployed are included only those groups of persons who come within their scope. "Overcrowded" in the usage of the Census reports means that the number of persons enumerated in a tenement is more than twice the number of rooms in it, a room being defined so as to exclude bath-room, scullery, etc.

It must be realised that the words describing statistical

18

totals or averages, such as *population, imports, tonnage of ships entered, average price, cost of living, occupied, wages, income, capital,* are technical terms, whose significance is always more definite than that usual in conversation or writing, and may have some essential difference from that in common usage. These terms are capable of exact definitions, which can only be ascertained from the original reports in which the totals are obtained, and it often happens that these reports leave serious ambiguities unsettled. The sections that follow in this chapter illustrate the examination of the raw material of investigations with a view to ascertaining the exact meaning of the totals obtained.

It is necessary in stating totals or averages to be as explicit as is possible without too much verbiage, and to give definitions which are too complex for a simple heading in juxtaposition to the table which contains them.
Explicitness in statement.

Thus in coal production we should not speak of " output per worker," but of " number of tons of coal brought to the surface in the week beginning January 25th, 1920, in the aggregate of the coal mines of Great Britain, divided by the average number of persons employed underground in that week," or if this is too complex all these points should be clear from the context or sub-headings or foot-notes, and an explanation should explain how the average number of persons employed was computed.

A percentage should never be given without a phrase showing on what it is measured. Thus if the price of some commodity was £80 at a previous date and is £100 now, the increase is 25 per cent. *of its earlier price* and 20 per cent. *of its present price.* If it now fell 25 per cent. *of its present price* it would reach £75 ; but if it fell 25 per cent. *of its earlier price,* it would return to £80. If wages are raised four times by 10 per cent. *of a standard,* starting at that standard, the wages are 100, 110, 120, 130, 140 per cent. of the standard ; but the increases in each period measured as percentages of the wage at the beginning of that period are respectively 10, 9·1, 8·3, 7·7 approximately.

A useful way of ensuring explicitness in a complex definition, of special importance in schemes of tabulation, can be illustrated as follows. In a table presented to the Income Tax Commission we find the sum
Attributes or characteristics.

£1,970,000,000 as the total of taxable income, explanations being given in introductory notes. The definition of this total may be exhibited thus :—

    A. Income.
    B. Known to the tax commissioners.
    C. As defined by the laws and instructions for assessment.
    D. Less allowances for wear and tear, etc.
    E. Of persons and corporations in the United Kingdom and of non-residents so far as they are subject to tax.
    F. Assessed for the fiscal year 1918–19.

Each of the six phrases expresses a *characteristic* or *attribute* possessed by every unit in the total, and the exact definition of these attributes leads to the definition of the total, and to the answer to the question " what has been counted ? "

We should generally include as characteristics, the fact of record (B), a date (F) and a place (E).

SECTION I.—THE POPULATION CENSUS.

The population census will provide good illustrations of the principles laid down in the last chapter, both because we shall
The census. be at first on familiar ground, since every one knows its scheme, purpose, and details, and because the form of inquiry used for the collection of the original data brings out very prominently the difficulties met with in detailed statistical investigations.

The first thing to be considered is the exact object for which the census is undertaken. It is for demographical pur-
Its object. poses ; to supply information as to the numbers and local distribution of the population, the numbers of each sex and age, their so-called civil condition (*i.e.*, whether single, married, or widowed), and their nationality. This is the minimum information necessary for administrative purposes. In addition to these facts there are very many others which the statesman and the economist wish to know about each member of the population, and the census form is the only means in England of collecting universal data ; the question as to which of these shall be investigated and
The choice of questions. which neglected, is decided more by expediency than on principle. Of these desiderata the follow-

ing may, be mentioned : the size and structure of the family, its position in the social scale, the economic position of its head ; the nature of employment of its members, the wage or income of each member and of the family as a whole, the rent and size of their house, their educational condition, the ages at which they commenced or retired from work, their migrations, their combination in religious or other bodies, and their infirmities. It is clear that some of this information must be dispensed with, if the form is not to be overcrowded, and if the tabulation is to be finished in any reasonable time ; and an examination of the general nature of the questions which can suitably be put will show how the necessary selection is made.

First, the questions must be those which the informant is able to answer. Now, if the questions were only to be put to educated and methodical persons, doubtless a *Ability* full account could be given of the family migra- *to answer.* tions and of the ages at which each member had been at work ; but the peculiarity of the census is that it is universal, and the questions must be such that the least educated and most unthrifty householder shall be able to answer ; in many cases such facts would have been unrecorded and forgotten.

Secondly, the questions must be perfectly definite, so that there can be no doubt as to what the right answer should be. The only answers which are of value to the statistician are " yes," " no," or a simple number, *Definiteness.* or a definite place or date or the use of a word that has a precise meaning. Adjectives and adverbs such as many, often, partly, etc., bear different numerical meanings to different people, and, though they may express fairly clearly the position of an individual, are nearly useless for tabulation,* which is their only purpose so far as the census is concerned. Thus the question as to education would have to be, not " state whether well, moderately, or badly educated," but " state at what age school was left," or " how many years at school ? " But even if such questions were not excluded by our first test, by the forgetfulness of the informant, the statements given would be of little practical value, and very often incorrect. An inquiry as to wage and income could not be made sufficiently definite without so many questions as to require a form to itself ; for wages, as we shall see when con-

---

* But see p. 121, *infra.*

sidering the Wage Census, require very careful definition, and many subsidiary questions must be put to get a proper estimate ; the simple query, " what is your weekly wage or annual income ? " would be answered on so many varying principles that the result would be valueless.

Thirdly, the questions must be such as will be answered truthfully and without bias. There is hardly a demand on the census form which would not be excluded, if *Veracity.* this rule was too rigorously enforced, as we shall see immediately. Perhaps the most difficult in this respect is the question, *Employer or employed ?* For though there are many cases in which a man is both employer and employed so that this question should be excluded by our second test, many persons consciously exaggerate their social importance by erroneously replying the former. Questions relating to social position must generally be excluded by this rule.

Fourthly, the questions must be those which will be answered willingly, and must therefore not be inquisitorial, or such as *Reluctance to* to raise apprehension of a change of law or an *answer.* imposition of taxes. Questions as to membership of trade unions, or of friendly societies, or as to insurance, would be thought inquisitorial. Many would refuse to state their incomes, holding it to be no one's concern but their own. Questions as to rent might be regarded as possibly leading to taxation. Questions as to religion are badly answered, as was shown in the evidence before the Census Committee of 1890,* and should be excluded in England by each of these four rules. Some persons do not know what their religion should be named, others would find the question indefinite, others would deliberately answer wrongly, and many not at all.

The questions on the census form † not excluded on one or other of these grounds are Nos. 1, 2, 3, 4, 14 and 15 ; these are fairly definite, and householders are generally able and willing to give correct answers to them. Question 5 may be inaccurately answered in cases of divorce, separation, or irregular unions. Questions 6, 7, 8, 9 were first introduced in 1911, and though there were many inaccuracies, the answers have given important new information. With regard to questions 10 and 11, there has always been difficulty in dis-

---

* *Report of Committee on the Census,* 1890 (C.—6071).
† Facing this page.

tinguishing between a classification according to the nature of the work a person is doing (*e.g.*, as a clerk or a carpenter) and a classification according to industries (*e.g.*, where the clerk and carpenter are employed by a textile firm); in 1911 questions 10 and 11 were devised with a view to making a double tabulation possible. Question 12 has failed to give complete information as to the status of a worker, and question 13 is inadequate for many cases. No provision is made for persons who follow two equally important occupations. Question 16 is not definite and leads to no important results A further discussion of the merits of some of these is to be found in the Report of the Committee already mentioned; * here it is only intended to indicate the general grounds of inclusion or exclusion.

So far we have not discussed the important question as to who should fill in the form. If, as in the English Census, it is to be filled in by the householder, the ques- Filling up of tions must be much simpler in matter and words the form. than if it is to be filled in by an official teller. In the latter case the form may be much more complicated, the questions more inquisitorial and such as might lead to indefinite answers on the part of ignorant people; for the teller would insist on an answer, be able to exclude those obviously wrong, and cross-question till the indefinite answers were so altered as to allow definite tabulation. In a great and complex undertaking like the Census, where many tellers must be impressed for a short period, their instructions and the general plan must be sufficiently simple; but as the extent of an inquiry contracts, the tellers can receive more complete instructions, and the information requisitioned may be more complex. This is of most importance in connection with columns 10–13.

The general shape and appearance of the sheet need attention. If the structure of the family is to be shown, the answers are best given on a single sheet, which must Shape of blank contain enough lines for the largest ordinary form. household, so that the trouble of fastening together of many couples may be avoided, and tabulation not be hindered. The spaces must contain plenty of room for answers in uneducated handwriting, without making the whole so large as not to lie

---

* See also the *Statistical Journal*, 1908, p. 496, and 1920, p. 134.

easily on a desk. The instructions must be distinct and visible, and placed in close connection with the answers; to further this, a skilful use may be made of capitals, italics, and different founts of type. On the form facing p. 22, those in use are roughly reproduced in miniature.

The form should always show for what purpose the figures are collected, and how they will be used, in order to enlist the
Purpose to be shown. support of the informant and allay misapprehension. The extent to which this should be done depends a good deal on whether the filling-up is compulsory, as in the population census, or voluntary, as in the wage census. In the case before us no preamble is necessary, since every one knows the main features of a census, and most are willing to further its objects; but it must be shown that the inquiry is sanctioned by Parliament, and that compliance is compulsory. This is done on the back, on the fold which is outside before the form is opened; and even though penalties are threatened against absence of or falsification of returns, the last sentence on the back and a statement on the front of the form guarantees the informant against injurious or personal use of his answers. Where information is voluntary, a careful letter should be printed and circulated with the form, persuading the informant to give his assistance.

While the main part of the form is filled in by the householder, other parts are filled in by the officials, and with very
Subsidiary information. little trouble a good deal of subsidiary information can be collected in this way. On the outside the Registration district and sub-district, enumerator's district and the postal address are written, from which the numbers can be tabulated for any of the areas required. The teller could also, as he took the form, enter the number of stories to a house, which is not done in the English Census, and other information as to the style of house and street might be endorsed. In a more intensive investigation, expert assistants could be trusted to come out of a house with an accurate knowledge of many interesting details.

We can now proceed to the individual criticism of the form in the light of the rules suggested above. In the first place,
Lines and columns. even the arrangement of columns is not perfect. To labourers who are not in the habit of writing at all, and who have (to judge from election posters) to be

instructed how to put their mark in the right place on a ballot
paper (many papers being destroyed simply through ignorance),
this arrangement of horizontal and vertical columns would be
confusing, and without help they would not gather at all what
they were to do. They would fill up more easily a paper in
which the answers were to follow the questions immediately :—

State your Name _____

State your Age _____

State your Sex _____

•Unmarried, Married, or Widowed _____

and so on.

This form, however, could only be used if a separate paper were
to be filled in for every individual, children and all, as is the
case in France.

    The first question, which for the general purpose of the
census should be the most definite of all, leaves some room for
doubt. What constitutes " passing the night "   <small>Criticism of the</small>
in the case of a night-watchman returning at   <small>questions.</small>
4 A.M., or of a printer at 2 A.M.? How is the householder
to know whether any of his establishment are returned else-
where? Since too many instructions only lead to confusion,
the tellers should be specially taught the answers to such
questions.

    The very meaning of the phrase " population of a district "
is open to much doubt. In France " la population de fait,"
which consists of all present in the given district   <small>Meaning of</small>
at the given moment, is distinguished from " la   <small>population.</small>
population de droit," which consists of all usually resident in the
district, including those temporarily absent, and excluding those
only momentarily present, and from " la population munici-
pale," which is " la population de droit," less prisoners, hospital
patients, scholars resident in schools, members of convents,
the army, and so on.* The English Census has counted only
" la population de fait." In the United States in 1890, we
find a " constitutional population," which excludes residents
in Indian Reservations, the Territories, and the District
of Columbia; the " general population," which includes in

_____

addition the Territories (except the Indian Reservation, Indian Territory, and Alaska); and the "total population," which includes all excluded in the former.* For 1910, the population as generally quoted is that of continental U.S.A., viz., forty-eight States (including Arizona and New Mexico, formerly territories) and the District of Columbia. We find also a total which includes Alaska, Hawaii, Porto Rico, and the army and navy abroad, and another total which adds to these the estimated population of the Philippines, Guam, Samoa, and the Panama Canal zone. For the apportionment of taxes the population of the District of Columbia and Indians are subtracted from the continental population. Notice that the Channel Islands and the Isle of Man are included in the English Census enumeration, but not in the total generally quoted. Also an account of soldiers and seamen at sea or abroad is given in a table, but they are not included in the total.

It is possible to find difficulties in filling up each of the columns, owing to ignorance or ambiguity. For illustration, consider how column 2 should be filled in in the case of a cousin who was a "paying guest," or a relation who was a visitor; for column 5, is a divorced person single or a widower, and what of a woman who is doubtful whether her husband is lost at sea?

It is well known that columns 3 and 4 are wrongly filled in for two reasons—one, that elderly people often do not Age. know their ages accurately and enter them to the nearest round number, so that the returns congregate at 40, 50, 60 : the error thus arising is eliminated by tabulation in the groups 35–45, 45–55 years, etc., and for more minute tabulation the groups 3–7, 8–12, 13–17, etc., are suggested : the other is that many women habitually enter their ages too low; in this case also the Registrar-General is able to deduce nearly correct totals.

It is to be noticed that, since the ages stated are those "last birthday," the age will on the average be given six months too low, and, in fact, the ages given as 17, e.g., should be scattered nearly uniformly over the months to the eighteenth year.

The most important criticisms of the census-schedule are to be made on columns 10–13. It will not be expedient here to go

---

* Willcox : *Area and Population of the United States at the XI. Census*, a book which gives a very useful criticism of the accuracy of the most elementary data of statistics.

DEFINITION OF UNIT · 27

into all the questions raised before the Committee on the Census as regards an industrial census. While there **Occupation.** can be little doubt that a thorough census of occupations would be best undertaken separately, and on somewhat different principles from the population census, it is certainly better, till opinion is ripe for so radical a change, to include in the present census the best questions we can as to occupations, than to omit them altogether in despair of accurate results. In any case, a census of occupations ought to be co-ordinated with the general population census, otherwise great difficulties of interpretation arise. Some of these may be seen in the attempt to reconcile the statistics of the number of persons employed in the *Report of the Census of Production* (Cd. 6320, pp. 8–10) with the statistics of persons occupied according to the Census of Population.

The objects aimed at, which we must always keep in mind when criticising special questions, are two : to find the number employed in each trade and industry, that is, so to say, to form vertical divisions; and to find the number in each kind or grade of employment (labourer, artisan, employer, etc., or smith's striker, carpenter, weaver, etc.) in horizontal divisions; so that the tabulation may give some such result as :—

TEXTILE INDUSTRIES.

	Cotton.	Wool.	Linen.	Totals
Employers -				
Managers -				
Clerks -				
Overlookers -				
Spinners -				
Weavers -				
Labourers -				
Children -				
Totals -				

The necessary minimum of information would be given by such answers as

> Legal—Solicitor—Managing clerk.
> Mining—Coal—Hewer.
> Metal-worker—Iron—Smith's striker.

Now the simple instruction, " State your occupation," would of course not lead to information of this sort. The coal-hewer would simply say miner; the clerk, managing clerk; the striker, very likely smith. To explain what is wanted and avoid mistakes, the informant is referred to the back of the form, half of which is devoted to instructions relating to these columns. These are lucid, carefully picked out with capitals and italics, comprehensive, brief and to the point. No one who wishes to fill in the form rightly, and is sufficiently educated to understand simple instructions, can easily go wrong. Yet it is probable that these instructions are in very many cases neither read nor followed; and this is very important in connection with the general study of blank forms of inquiry. Forms issued to people uninterested in the object in view will generally be filled in with the least possible expenditure of time and intelligence. Hence two courses are open : to reduce the question to the simplest possible form, and make the best of the result; or not to allow the informants to write in their own answers, but to take them *vivâ voce* by means of a teller, who has mastered the instructions, and has the necessary legal force behind him to compel information. The latter course entails time and expense.

The result of the present system of inquiry, combined with a faulty method of tabulation, which it to some extent makes necessary, is that we have no reliable census of occupations for the United Kingdom. The present figures break down both from faulty data and from insufficient tabulation directly we attempt to make some of the important calculations depending on them.

An attempt was made in 1891 to correct to some extent our ignorance of the relative numbers of unskilled and skilled labourers, employers and employed, by the question now in column 12. The headings were not a model of clearness; there was not the ordinary imperative

" state " or " write," nor was one told on the front of the form whether to write Yes or No or to make a mark in the appropriate column, nor is the distinction between the three headings a perfectly definite one; but still one was hardly prepared for the following statement in the report : *—

" In numerous instances, no cross at all was made ; in many others, crosses were made in two or even all three columns, and, even when only one cross was made, there were often very strong reasons for believing that it has been made in the wrong column. Oftentimes this use of the wrong column can scarcely have been other than intentional ; being dictated by the foolish but very common desire of persons to magnify the importance of their occupational condition. This desire must have led many subordinates to return themselves as employers rather than as employed, for it is only on this supposition that we can account for the otherwise unintelligible fact that, under several headings, there are actually, according to the returns, more employers than employed, more masters than men. . . . We hold (these returns) to be excessively untrustworthy, and shall make no use whatsoever of them in our remarks."

The questions have, however, continued to be inserted and the numbers tabulated, and statisticians have used the results with a certain confidence.

This attempt and its results are of the greatest importance to all who try to draw up forms of inquiry.

Before leaving the subject, it should be mentioned in passing that we cannot deduce directly from our census the number of persons dependent on a particular trade for their living ; that is to say, the number of employers, their families (not otherwise returned) and domestic servants, and the number of employés and their dependent families. This, the most important total for estimating the relative importance of different trades of the country, is not tabulated, though such tabulation has been found possible in other countries, and we are dependent on the estimates of statisticians for such totals.†

To see how the information given by the answers on the census schedule can be worked up into detailed specific numbers, it is only necessary to look at the diagram and

---

* *General Report on the Census of* 1891, p. 36 (C.—7222 of 1893).
† See Booth in *Statistical Journal*, vol. xlix.

table prefixed to each of the sections relating to special trades in Mr. Booth's *Life and Labour of the People* (*e.g.*, vol. v., p. 46).*

Statisticians have generally to work with material provided for them; their first task is to understand exactly the defini-tions under which the data were obtained and the limitations of the tables published. In skilled hands quite faulty compila-tions have often been found to yield accurate results of great interest.

SECTION 2.—THE WAGE CENSUS.

The main differences in method between the wage census, as taken in 1886 and 1906, and the general population census are—(1) That the filling up the forms in the wage census was voluntary; (2) that their correct filling up required a higher degree of intelligence and education. As before, we must consider first the object which the wage census was intended to fulfil: it was to describe the earnings of the people of the United Kingdom, to compare the rates of wages trade by trade, and to find the relative numbers earning at each rate. What is the best quantity to measure with this object in view? As a preliminary question, should we take the day, week, or year as the unit of time? Clearly we shall not be able to compute weekly wages if we only obtain daily, for the week's work varies from four to seven days in different occupations. The week's wage is a more definite quantity; but the simple comparison of weekly wages in different trades will be decep-tive, because most trades are busier at one season of the year than at another, and in many the difference between season and season is very great; in any particular week, then, we may be comparing the best season of one industry with the worst of another. To avoid this error, and because we do not know how many full weeks' wages are obtained in a year, except in a few non-intermittent trades, it would seem best to take the year as unit; but the direct calculation of an individual's annual earnings is practically impossible. The employer is not acquainted with this sum, for in large establishments the hands are continually changing, and one man will be paid by two or

The object.

The unit of time.

---
* See p. 57, *infra*.

more masters in the same year; and even in a factory with a
nearly constant personnel, the weekly amounts paid to indi-
viduals are not in general so tabulated as to be easily summed,
and the working out of the totals would require a prohibitive
amount of clerical labour.   If we turn to the workman, on the
other hand, we 'shall find in the majority of cases that no
accurate account has been kept of earnings through the year,
and it would only be by careful individual examination, im-
practicable on any large scale, that an estimate could be made;
in many cases the men, even if willing, would be quite unable
to give a connected account of their earnings during the past
twelve months.

It seems clear that we must adopt a smaller unit, and since
most wages are paid weekly, a week is the most natural one.
The subsidiary questions which will lead best to an estimate
of annual earnings will be discussed below.   The answer to
the first question, as to the best quantity to investigate, is
indirect; the only individual measurements we can obtain
directly are the week's wages, but these may be supplemented
by estimates *en masse*.

Next, who possess the information we require?   Clearly
both employers and employed, and in an ideal census the
answers would be obtained from both groups;   Employers and
but considerations of simplicity, cheapness, and   employed as
accuracy are all in favour of applying to em-   informants.
ployers alone.

If employés were to be interrogated the procedure would be
as follows.   Draw up a form on the analogy of the census form,
describe very briefly the purpose of inquiry, add a short series
of concise, lucid, simple questions in suitable type and with
careful spacing, such as will lead to the minimum information
required; let these forms be left to be called for, and when
collected, let the tellers have time and opportunity to examine
and correct them.   It is clear that this method would entail an
even more expensive organisation than the population census,
and as the result of experiment it may be doubted whether the
maximum of accurate information that could be thus obtained
would come up to the minimum that would be of use.   A
partial inquiry can, however, be carried out by means of trade
unions, as was the case in the census of Railway Wages under-
taken by the A.S.R.S. in 1908.

The method of inquiry among employers was as follows : Suitable blank forms and an explanatory letter were sent by post to all employers, whose addresses could be found, in the industries selected for investigation, and the answers were returned to the central office by post. This is far simpler and cheaper than the suggested scheme for inquiry among workmen, requiring far fewer forms and only a small staff of clerks. With business men it is a simpler matter to post the return when completed than to keep it for collection by hand. Since there is no personal intercourse over the matter it is especially necessary that the questions should be lucid, for the additional correspondence necessary to rectify errors is a source of worry at both ends. A copy of one of these forms used in 1886, abridged only in the number of subdivisions, is subjoined here and on the following page.

WAGE CENSUS.

RETURN OF THE RATES OF WAGES PAID IN SILK MANUFACTURES.

*Name of Factory or Firm*_____

*Address*_____

_____

Note.—It is requested that the salaries of clerks and managers may be excluded. The return is of wages of working men only.

Numbers employed on_____ 1886 -   - No. _____

Amount paid in Wages in the year 1885 -   - £__ _____

Highest weekly amount paid in 1885   £_____ Date ____•___

Number of Hands paid in that week   -   - No._____

Lowest weekly amount paid in 1885   £_____ Date _____

Number of Hands paid in that week   -   - No. _____

State the present average rate of pay for overtime : that is, whether overtime is reckoned as time and a quarter or time and a half, &c., or in what way reckoned _____

State whether overtime is at present being worked, and how much ; or whether less than full time, and how much less_____

CURRENT RATES OF WAGES AND HOURS OF LABOUR PER WEEK of Persons employed in each Branch of the Silk Manufactures on _____ 1886.

DESCRIPTION OF OCCUPATION. N.B.—It is requested that this list of occupations may be revised where necessary	CURRENT RATES OF WAGES PAID and NUMBER OF HOURS OF LABOUR PER WEEK when in full work, but exclusive of Overtime. Note.—State the Number of Hours of Labour per Week, whether the Workers were paid by Time or Piece-work, and if paid by Piece-work give the amount earned in a week, exclusive of Overtime.											
	MALES.						FEMALES.					
	MEN.			LADS & BOYS.			WOMEN. 18 years and upwards.			GIRLS. Under 18 years.		
	Number Employed.	Rates of Wages.	Hours of Labour.	Number Employed.	Rates of Wages.	Hours of Labour.	Number Employed.	Rates of Wages.	Hours of Labour.	Number Employed.	Rates of Wages.	Hours of Labour.
Silk Throwing—												
Parters — { Time / Piece												
Winders — { Time / Piece												
Cleaners — { Time / Piece												
Spinners — { Time / Piece												
Doublers — { Time / Piece												
&c.												
Silk Spinning—												
Openers and Sorters { Time / Piece												
Boilers — { Time / Piece												
Dressers — { Time / Piece												
Preparers and Carders { Time / Piece												
&c.												
Silk Weaving—												
Winders — { Time / Piece												
Warpers — { Time / Piece												
Warp Pickers or Clearers { Time / Piece												
Doublers — { Time / Piece												
Fillers — { Time / Piece												
&c.												

D*

The measurement of the annual earnings of groups of workpeople was one of the ultimate objects of the inquiry. *Annual earnings.* Annual earnings are composed of many different items, of which the following are the most important : Ordinary weekly wages, pay for overtime, special payment for special work (*e.g.*, of builders if sent to a distance), or at special seasons (such as the harvest) ; and payments not in cash, such as free or reduced house-rent, free or cheap coal, and special goods at cheap or wholesale prices (such as cloth in textile factories, or potatoes for agricultural labourers).

When payment in kind is at all general or important, it is generally better to proceed on a different method entirely, *e.g.*, that followed by the Agricultural Sub-Commissioners of the Labour Commission. When it consists of only one simple item, such as a house rent-free, it can form the subject of an additional question on a form similar to that on p. 32. In the silk industry this does not occur ; but this discussion shows the necessity of preliminary knowledge on the part of the investigator before the right form of inquiry can be drawn up. .

We have left for consideration the weekly wage, and overtime and special payments, the last two of which can be grouped together. The ordinary weekly wage is a sufficiently general and definable quantity in most subdivisions of most industries. A foreman could generally state how much is earned in an ordinary full week for each of the hands under him. In many cases there is an hourly or weekly sum regulated by a trade union, as in the building trades. In others, as in the cotton industry, piece-rates are so regulated as to bring out a definite sum for the week's work graduated in relation to the difficulty of the task; in general, a very rapid survey of the wage-book will show what the worker in each subdivision will make on an average. Thus the average weekly wage in an ordinary full week can be found with considerable accuracy, but this takes us only part of the way in the calculation of annual earnings ; we need to know in addition to this how many full weeks are made in the year. It is the method by which this is attempted on the printed form that is open to most criticism. The questions used are on p. 32, and afford a good example of the general difference between the quæsita and the data which are attainable. The quæsitum is : To how many full weeks' wage are the annual earnings equivalent allowing for slack weeks

and overtime? The first crucial question to decide is: Are we
to allow for an average loss of time, say two weeks in the year,
through sickness, or are we to allow only for time   Quesita and
lost through failure of work? Since sickness is   data.
an individual not a general misfortune, it will be better to
exclude it if possible. Now overtime in one season, especially
if its wages are on " time-and-a-quarter " or " time-and-a-
half " basis, very quickly tends to balance slack time at another
season, though it may be supposed that it is rarely the case
that more than the normal week's wage is averaged through
the year. Thus it will be logical as well as simple to estimate
the year's earnings as so many normal weeks' wages. For
example, if we found that two weeks were lost through sick-
ness and three through the mill stopping, and that overtime
in one busy month had added wages equivalent to two normal
weeks, we should have forty-nine weeks' full wage. The
figures which will give this result will be the total sum paid in
wages in the factory in the year divided by the aggregate
normal week's wage of the people dependent on the factory,
supposed all at work. Thus, if 1200 hands (men, women, and
children) would, if all at work, make £1000 in a normal week,
and this was the average number dependent on the particular
mill, and if £48,000 was paid in the year in wages, annual earn-
ings would be equivalent to forty-eight normal weeks, and
earnings would average £40. Now the total paid in wages is
generally kept separate in business accounts, but the number
dependent on the mill for work is often not known accurately;
for the personnel of a large establishment is subject to continual
change, and the manager would not know whether a person who
left went to another mill or got no work. The total number
of all who had worked there during the year would be too great
for this purpose, and the number at work in a normal week
too small. The number open, perhaps, to least objection is
the number at work in the busiest week of the year; for those
absent except through sickness when trade is busy cannot be
said to be dependent on the factory, but if not at work else-
where are among the permanent unemployed; very few work-
people indeed will be taking their holiday at a busy time, and
it may reasonably be supposed that all the factories in the
same industry will have their busy and slack seasons at nearly
the same time. The answers then to the printed questions—

Total paid in year, and number of hands in busiest week—tell us all we need to know, if we may make this assumption; for then the total sum paid as wages in the year, divided by the maximum number employed in the busiest week, gives the average annual earnings. To find the equivalent number of normal weeks, multiply the maximum number employed by the average wage found on the second page of the form, so that the product shows the aggregate weekly wage if all were employed, and divide the total paid in the year by this product.

The process may be illustrated by comparing the data obtained in the more recent census with that in 1906, when more information was obtained.

In 1906 some of the particulars obtained were as follows. The Cotton Industry of the United Kingdom is taken as an Annual earnings example and the figures relate only to those and weekly firms which made returns. [See Cd. 4545, pp. wages. xxv–xxxvii, 3, 17, 20–28, and (for the blank schedule issued) 242-4.]

T = Total wages in 1906 = £10,195,229.

W = Average of 12 weekly statements * of aggregate wages = £204,173.

N = Average of 12 weekly statements of aggregate numbers = 212,503.

M = Greatest aggregate recorded among N = 213,472.

A_4 = Average earnings of all employed in particular week = 19·43s.

A_I = Average earnings of those employed in particular week who worked neither overtime nor broken time = 19s. 7d.

Hence we have

$$A = \frac{W}{N} = \text{average earnings in the 12 selected weeks} = 19\cdot21s.$$

$$E_a = \frac{T}{N} = \text{average annual earnings of the average number} = £47\cdot98.$$

$$n_1 = \frac{E_a}{A} = \text{number of weeks' average earnings obtained in the year} = 49\cdot95.$$

The difference between 52 and n_1, i.e., 2·05 weeks, is attribut-

able to·holidays, which range from 8 to 15 working·days, but includes also stoppages of the factories from any cause.

$E_m = \dfrac{T}{M} = £47\cdot76$ = average annual earnings of the maximum number, which is taken as the number dependent on the factories, the variation·being due to unemployment.

$n_2 = \dfrac{E_m}{A} = 49\cdot73$ = number of weeks' average earnings of this maximum.

$n_1 - n_2 = 0\cdot21$ = possible estimate of weeks lost by unemployment in 1906.

To this should be added an estimate for the number unemployed in the maximum week.

$\dfrac{A_f}{A_e} = 1\cdot008$. In this case broken time exceeds overtime on the whole, so that earnings are 0·8 per cent. below those obtained simply by full-time work.

Applying this percentage to n_2, we obtain 49·34 as the number of weeks in the year in which full-time earnings could be obtained by the maximum number recorded as employed.

•In 1886 the corresponding totals recorded were T = £3,148,566 for the year 1885. A_f, ordinary wages in a normal week in 1886 = 15·2s. M, the greatest number recorded in 1885 = 87,887. Hence $E_m = \dfrac{T}{M} = £35\cdot8$, average annual earnings of maximum number; and if we can take A_f as the same as A (the average weekly earnings in the year), $n_2 = E_m \div A_f = 47\cdot1$, the number of weeks' earnings of this maximum. Here we cannot compare A_f and A_e for want of data.•

The method is evidently open to criticism from several points of view, and is here given rather to illustrate the nature of the problem and of the data which may help to solve it, than as a complete statement of the relation of normal wages to annual earnings.

In·addition to lost time due to holidays and to complete unemployment in the maximum week, there is lost time·due to sickness, of which an estimate * is an average of 1·7 weeks.

In the corresponding French wage census, of which the

---

* See *Division of the Product of Industry*, 1919, by the present author, p. 30 and Dr. Snow in the *Statistical Journal*, 1912–13, p. 477.

results were published in 1898,* an estimate of the number of days' work obtained in the year is formed on a different basis. The data collected were—(1) The variation each month of the personnel in each industry, which is found to average 4 per cent. **The French method.** for the year—that is, for each 100 employed, 96 are found who have been in the same establishment for as much as twelve months : (2) The differences between the maximum and minimum numbers employed in each establishment month by month during the course of a year, which are found to average 19 per cent. of the (? average) personnel. From this we may perhaps draw the conclusion hat, on an average, half this number, at least, are in general out of work : (3) The number of different persons who have been employed in each establishment at one time or other in the year; this is found to be 140 for each 100 permanently employed, from which the legitimate conclusion is that the average number of unemployed is not so much as 40 in 140, *i.e.*, 28 per cent. These two percentages, 9 per cent. and 28 per cent., are taken to be the inferior and superior limits of average lack of work. This information is more detailed and perhaps more reliable than that on which the method, used above for the English figures, is based. Data obtained from syndicates of French workmen indicate about 20 per cent. as the average want of work; the English figures obtained by the method described above from the whole wage census yield about 12 per cent. in 1886.

This somewhat lengthy discussion on the few questions included on the first page of the form is a good illustration of the necessity of considerable preliminary study before a blank form can properly be drawn up. Space does not allow a detailed criticism of the rest of the form, but it should be mentioned that the questions relating to individual wages in 1886 were not sufficiently detailed. Thus under "Spinners, piece" (see schedule, p. 33) in each factory the earnings given would be an average for all employed, so that the earnings of individuals were not recorded, and the general distribution of earnings could only be given approximately. In 1906 the instruction was "Those earning the same amount may be grouped together; otherwise each entry should represent only

---

* *Salaires et Durée du Travail*, 1897, pp. 15, 16. *

one person," and the actual variation in each occupation and industry could be shown.

A careful comparison of the two schedules is recommended, for it will throw light on many of the difficulties experienced in preparing questionnaires.

SECTION 3.—EXAMPLE OF AN UNOFFICIAL INVESTIGATION.

Investigations without official authority do not differ essentially from those conducted by authority if (as in the Wage Census) there is no compulsion to answer; but they are generally more limited in their scope, for want of organisation or funds, and are at the same time freer to employ the method of samples (which is discussed below, Part II, Chapter II.) and induced to do so in order to cover an adequate field.

As an example, we may take the investigations relating to the economic condition of the working-classes in certain towns, whose results are published in *Livelihood and* Working-class *Poverty.** The problem to be considered was not conditions. precisely defined beforehand; in brief, the intention was to obtain what information it should be found practicable to get, as to the number of earners and dependents in working-class families, their earnings and their needs, and to tabulate those parts of it which after criticism were believed to rest on trustworthy answers.

It is generally the case in such investigations that it is necessary to obtain the information personally, since people are not willing to fill in and return questionnaires unless there is some strong inducement (*e.g.*, obtaining sugar) to do so. Consequently the forms used need contain few instructions, the investigators being specially selected and prepared for the work. It was found advantageous to use cards rather than paper schedules, and a facsimile is given on p. 40.

It had, as always, to be considered what facts were actually known by the householder or his wife, and what were likely to be communicated to a tactful and persistent The blank card inquirer Once the wife is engaged in conversation, there is no difficulty in eliciting information as to the inhabitants of the household, the age of those under twenty, and the occupations (and generally the employers) of those

---

* Published for the Ratan Tata Foundation.   G. Bell & Sons, 1915.

at work; the rent and type of house can also be easily recorded (and checked if necessary). This information by itself led to valuable tables showing the constitution and earning strength of the families, and the enormous variation and absence of

File No.													Enq'ry No.					SEX OVER

EARNINGS. full time
EARNINGS. last week
HOURS. last full week time
Employer
Occupation
WAGE EARNERS. Relationship Age
Non-Wage Earners: Sex Age Relationship
Other Sources of Income.
Housing: Rent, Kind of House, No. of Persons, No. of Rooms, Garden

any standard type in these, of a kind that have not been compiled in the census or any other official investigation.

The difficulties were found in assessing family income. The wife often does not know the husband's or the elder

children's earnings, and information is not readily given in very many cases. Where it was given, it could be verified in selected cases by inquiry from the employers, and where tests were made it was found that there was no bias in the direction of either overstatement or understatement. If, as was generally the case, the occupation was correctly stated, it was possible to estimate with fair accuracy the normal week's wage from the known standard in the town. The distinction on the card between " last week's " and " full time " earnings was made because the first was capable of a definite answer, and the second was often an estimate. The investigator, having both statements, would be able to find the reason for any difference, and to establish the second (which was the only one used in tabulation) more definitely than if it stood alone. It was found necessary to tabulate only the conditions which would exist if a full week's work were done, and to leave aside questions of sickness and unemployment. The answers to the question as to " other sources of income " were certain to be imperfect; but, so far as they went, they showed the means of livelihood of some families whose wages were evidently insufficient, and since they erred only by omission they gave some positive information. The majority of working-class families have only a negligible amount of income-yielding property, and the main source of such income, the ownership of the house inhabited or of other houses, was generally reported.

The estimates of earnings were not believed to be sufficiently accurate to lead to a table showing the numbers with various annual incomes, but they were adequate for the main purpose for which they were used. This purpose was to find out what proportion of the families had an income (apart from charity) to bring them above a certain standard, such as Mr. Rowntree's minimum standard as calculated by him in *Poverty*. In the great majority of cases there was no doubt from the constitution by age and sex of and the number of dependents in the family and the nature of the man's work on which side of the line the household stood. In the doubtful cases (which were kept apart in the tables) advantage was taken of all the points noted by the investigator (including non-numerical statements written on the back of the card which was reserved for this purpose), and a reasonable judgment could generally be made.

Relation to minimum standard of living.

The card was not shown to the informant, but was filled in immediately after the conversation. The identity of the household was only preserved by the inquiry number. A file number was written in to preserve an order after the first process of sorting. Each card was criticised, and the numbers needed for tabulation were computed, and these and abbreviations showing the constitution of the family (such as m., s. : w., sc., sc., in., where a man and his son were earning and his wife, two school-children, and an infant were dependent) were written in the small spaces under the words " File No." The entries for the tables were then obtained by dealing the cards into appropriate packs and then counting them; this process is rapid, but needs continual careful verification.

The scope of each inquiry (e.g., the working-class of Northampton) needed careful definition. It had to be decided

Definitions.

whether the town from an economic point of view coincided with the administrative borough, and, if not, outlying houses must have been definitely included or excluded. Next it was necessary to get an accurate list of all the houses in the district and apply the method of selection by sample to this list ; the inquiry actually dealt with whatever was contained in the list used, and the list gives the definition of its scope. There is no accepted definition of the " working-class," and that actually used was in fact determined during the handling of the cards. As a preliminary, all the houses at first selected which were above a certain rental or whose tenants were contained in a directory of principal residents were excluded. Of those visited all were excluded in which the principal earner was a clerk, teacher, or manager. For others, such as shop assistants, commission agents, publicans, small shopkeepers, decisions had to be made and recorded as the various cases arose. The final definition of the working-class households, as understood in the inquiry, was then by delimitation, and if given in full would be somewhat as follows : all households where the rent was less than 12s. weekly, in which the principal earner was not a clerk, teacher, etc. etc. Such a process of forming the definition during tabulation is of necessity quite common; the decisions should be quite clearly shown in the report, and emphasis should there be laid on the treatment of marginal cases.

SECTION 4.—STATISTICS OF ENGLAND'S FOREIGN TRADE.

The original schedules which lead to many other statistics are interesting, but limits of space must restrict us to one more typical inquiry, that which leads to our statistics of foreign trade.

In the population census the filling in of the form is compulsory and done by the householder; in the wage census the answers were voluntary and given once and for all by the employer; in the various inquiries undertaken by the Labour Department the answers are voluntary, but in many cases periodic, so as to become quasi-official. The method of collection of import and export statistics is a blend of all these. There are three classes of persons who know the facts in question—the sender of the goods, the *The Informants.* custom-house official through whose hands they pass, and the recipient or his agent. Circumstances decide that, in the case of exports from the United Kingdom, the exporter or his agent sends an account of the quantity and value and place of destination, etc., of goods despatched to the Statistical Department of Customs; that, in the case of imports, the receiving-agent hands over an account of goods to be landed to the custom-house officials, who verify the account, roughly if the goods are duty free, carefully if they are liable to duty; and that, in the case of transhipment, the goods are treated in the same way as imports at the port of landing, and to some extent verified at the port of embarkation.

The blank forms, being verified by officials as part of their duty, or having been filled in by agents thoroughly used to the task, need no covering letter, and may be made as complicated as necessary; no questions are inserted but only blank tables. An examination of the forms in use will show what are included as exports and imports in the Board of Trade totals, and what is the total amount of information available for tabulation.*

The quantities we wish to measure in this investigation are : the volume or weight and value of all goods which have an exchange value, which leave our shores or reach *The quantity* them from without, subdivided as regards classes *and data.* of commodities and countries of destination or origin; the

---

* The following paragraphs do not take cognisance of any changes that may have taken place since 1914.

values being those at the times of loading or unloading. The quantities we can measure are sharply distinct from these, being the records of values and volumes which reach the Board of Trade. We should therefore examine the forms to decide—(1) What part of imports and exports are recorded; (2) whether the values are correctly given, (3) the quantities accurately registered, (4) the commodities accurately defined, (5) the countries of origin and destination accurately distinguished in the returns.

On reaching port the ship's master has to send in an Examples of information. account, of which an abridged specimen is given on p. 45 :—

The goods for quick transit are passed at once, and a special form is sent to the Customs Establishment similar in character Dutiable goods. to that on p. 46. The remaining goods are treated either as dutiable or as duty-free articles. In the list before us, ten cases of wine are entered for home use, and an account is sent into the Statistical Office; sixty cases are warehoused and another account (as to quality, quantity, and value) is sent in; the whole are registered as imports. Twenty of the warehoused cases are removed to another port and re-exported; an account is sent, and they are entered as exports of foreign goods. Twenty are put on board ship as stores at the port of entry, and ten more removed to another port for the same purpose, and of this the central office receives an account; the remainder are removed to another warehouse, still in bond, and on leaving that will be treated in one of the four ways just mentioned. Other dutiable articles are treated in the same way.

Goods not sufficiently described or not answering to their description are opened, their contents entered on a " bill of Examination of goods. sight," and an account sent in. Private effects are separately examined, being described on a " sufferance " form; if they are *bona-fide* personal goods no record is kept of them, except in the case of dutiable goods, which are treated as ordinary imports. If the dutiable goods are concealed, either among private effects or merchandise, and forfeited, they are not reckoned as imports.

Bullion is entered on a separate form and kept distinct throughout the accounts.

The duty-free goods, if for transhipment at another port,

are sent, there under seal, and barely examined; they are treated at the central office in the same way as dutiable transfer goods. The remaining free goods, **Free goods.**

If Sailing Vessel
or Steamer , STEAMER.

Official No.
No. of Register,
Date of Registry,

No. 1.
Port of X.

REPORT No. 980.*

Ship's Name.	Tonnage.	British or Foreign. If British, Port of Registry; if Foreign, Country to which she belongs.	Number of Crew.		Name of Master, and whether a British or Foreign Subject.	Port or Place from which arrived.
			British Seamen.	Foreign Seamen.		
Marianne.	700	BRITISH. Total..	12	—	H. Hind.	Havre, France.

CARGO.

1. Name or Names of Places where laden in order of time.	2. Marks	3. Nos.	4. Packages and Description of Goods, Particulars of Goods stowed loose, and General Denomination of Contents of each Package of Tobacco, Cigars, or Snuff intended to be imported at this Port.	5. Particulars of Packages and Goods (if any) for any other Port in the United Kingdom.	6. Goods (if any) to be Transhipped or to remain on Board for Exportation.	7. Name of Consignee.
Havre, France.	Paris	s to	London.—600 pkgs. 68 pkgs. Merchandise.	Fruit and Perishables.		Smith.
	COK	1392				
	AE	495/6				
	KG	340/9				
	FOT	1/50				
	AJ	3/6				
	CK	1				
	AC	10				
	KL	40	70 cases Wine.			,,
If any wreck fallen in with or picked up, to be stated.	ACD	20				
	WD	166	5 cases Woollens in transit to Liverpool.			,,
	O&D	1	1 case Brandy.			,,

STORES.

Surplus Stores remaining on board, viz. { 3 lb. Cigars.
                                          4 lb. Tobacco.
Number of Alien Passengers (if any) - Nil.
Pilot's Names   -   -   -   -   -   ——
At what Station Ship lying -   -   - South Quay.
Agent's Name and Address -   -   - C. J. C.

I declare that the above is a just report of my Ship and of her Lading, and that the Particulars therein inserted are true to the best of my knowledge, and that I have not broken Bulk or delivered any Goods out of my said Ship since her departure from Havre, the last Foreign Place of Loading.

(Signed)   H. HIND, Master.

Signed and declared this 13th day of October 1896
   In presence of
      (Countersigned) *
_____ pro- Collector.

* i.e., 980th ship at X. since 1st January.

which in general form the bulk of the cargo, are entered on such a form as follows, which is worth notice, for it is a specimen of the rough material from which our foreign trade figures are evaluated.

ENTRY FOR FREE GOODS.*

This space is for the use of the Officers of Customs.	Port_____ Dock or Station_____ Importer's Name_____ (No._____)

Examination.	Ship's Name. Marianne.	Master's Name H. Hind.	Rotation No. 980.	Date of Report 13/10/96.	Port or Place whence Havre, France.

Marks and Nos.	No. of Packages and Description of Goods, in accordance with the Official Import List.	Quantity.	Value, £.
COK 1392	One Goods Manuf. N.O.E. Billiard Cue Tips - - - -	...	28
AF 495/6	Two Leather Shoes - - -	10 doz. prs.	58
KG 340/9	Ten Cotton Manuf. Trimmings -	...	140
	Embroideries	...	280
	Piece Goods, not Muslins - -	300 yds.	8
FOT 1/10	Ten Gloves of Leather • •	11,240 doz. pr.	12,316
,, 11/5	Five Silk Broad Stuffs • •	...	10,400
,, 16/20	Five Works of Art—		
	Plaster Casts • • •	...	380
	Statuary - - • •	...	1,280
	Pictures by Hand • •	3	10,200
,, 21/5	Five Books Bound - - -	4 cwt.	300
,, 26/30	Five Bronze Manuf. Ornaments -	3 cwt.	38
,, 31/5	Five Metal Manuf. Ornamental Brass-headed Nails - -	4 cwt.	24
,, 36/40	Five Silk Manuf. Dresses, Mantles, Trimmings - • • •	...	1,816
,, 41/50	Ten Goods Manuf. N.O.E.—		
	Fancy Goods - • •	...	110
	Horseless Carriage • •	...	160
	Brushes - • • •	...	78
	Glue - - • • •	...	110
	Billiard Chalk • • •	...	12
	Hardware - • • •	...	116
AJ 3/6	Four Stationery Ink - • -	...	48
CK 1	One Iron and Steel Manuf. Machinery, British, returned	3 cwt.	24

   I enter the above goods as free of duty, and declare the above particulars to be true.

   Dated this 13th day of October 1896.

                (Signed)   **J. Jones,**
                                  Importer or his Agent.

* In 1904 this form was altered so as to distinguish between " place of *shipment* of goods," which phrase replaced " whence " in the last heading, and " place whence goods *consigned* " which is now the heading of an additional column.

The information so received is usually accepted at the central office without inquiry. It frequently happens, however, that the form is not properly filled in by the agent, the values often being omitted. When this is so, it is the Verification of data. duty of the clerk at the port of entry to require the agent to complete the forms, if imperfect, and to test the values by current price lists with which he is provided. When there is a palpable error or omission in the form, or when the price appears out of the common, a query is sent from the central office to the port : e.g., with reference to such a form as that just given, the following correspondence might arise :—

1. Pictures by hand, £10,200. Explain high value. *Answer.*—Correct ; invoice was seen ; pictures by Millet.

2. Books bound : is weight or value incorrect ? *Answer.*— Both correct ; advice seen ; old and valuable books.

3. Goods entered as " goods manufactured, chip plaiting " : explain nature, and state if description is correct. *Answer.*— Correct ; wood shaving plaited and occasionally mingled with horse-hair, etc.

4. Potatoes, 40 cwt., £62. Weight or value ? *Answer.*— Value correct. Weight should be 400 cwt.

Thus any unusual entries are liable to be checked and verified.

In the case of goods not easily valued, or of miscellaneous goods not easily tabulated, errors must arise in this way ; and another error may enter if an agent or clerk, who Possibility of errors. does not wish to receive too many queries from headquarters, enters at ordinary rates goods of exceptional value ; but when staple commodities and large quantities are involved, all the persons concerned will be familiar with the forms they have to fill, the prices will be known, and so in important cases errors will be at a minimum. The import total values, therefore, are the sum of many quantities of various degrees of accuracy, and it is not difficult when looking through the list of items in the annual report to see which are specially liable to error. Such commodities as old books, works of art, goods where sale depends on the fluctuations of fashion, racehorses, and so on, have values varying from day to day, and their exact value in the balance of imports and exports cannot be determined.

In the case of goods consigned for sale, a class which includes the great part of the imports of wool no value can be named by the agent. The goods are then valued at current market prices, and in the case of wool at the prices realised at the next wool-sales. There is always a possibility of error here, since the current prices may not be exactly obtained for a particular consignment; and there is apparently permanent overvaluation of wool, since the price at the sales is presumably the price of wool landed and warehoused, while the value for import records should exclude the cost of unloading and moving.

The quantities and values of exported goods are filled in by the shipper or agent, and the papers sent through the

Exports.   Custom House officials or directly to the central office within six days of the ship's clearing. The specification given on p. 49 is an abridgment of the form used :—

The forms for British and Irish goods are distinct from those for foreign, free and duty-paid, goods; and there are distinct export forms for transhipments, which have already been registered as imports. In these cases the specification and quantities are likely to be correct, but there are causes which may falsify the values. If they are to be subject to an *ad valorem* duty, they may be undervalued; if they are adulterated goods, masquerading as genuine, they may be overvalued. It seems hardly possible to estimate these errors.

We are now in a position to define imports and exports according to their meaning in the Board of Trade Returns;

Definition of official imports and exports.   as, for instance, when for 1913 the value of imports is stated as £769,000,000, and of exports as £635,000,000, of which £110,000,000 are re-exports of foreign or colonial goods. In the following statement the details already shown are supplemented from the definitions given in recent years in the introduction to the *Annual Statement of the Trade of the United Kingdom.*

Under imports are included all goods landed through the custom-houses, including goods immediately shipped as stores or returned from customers unused, with the following exceptions : (*a*) fish of British taking landed in British ships arriving direct from the fishing grounds, goods directly imported by ambassadors and ministers accredited to this kingdom, *old* vessels bought from foreigners; and (*b*) sacks, cases, etc.,

*SPECIFICATION FOR BRITISH AND IRISH GOODS ONLY.†

Port of X. Ship Name—"Marianne." H. Hind, Master, for Havre.

*The Specification of Goods exported must be delivered to the proper Officers of Customs within six days from the time of the final clearance of the ship, as required by the Customs Laws.

Marks.	Numbers.	No. and Description of Packages.	Quantity and Description of British and Irish Goods, in accordance with the requirements of the Official Export List.	Value.
KCL	641/2	Two bales	Woollen Heavy Cloths. 1,400 yards - - -	£340
CKD	140/1	Two cases	Steel Manuf. Blades. 3 cwt. - - -	24
RMO	10/12	Three crates	Steel and Iron Manuf. Threshing Machine. One -	380
CL	140	One case	Cotton Piece Goods, Printed and Dressed. 1,800 yards -	50
			Total - -	£794

I declare that the particulars set forth above are correctly stated.

(Signed) ————— *Exporter or Agent.*

Address—Harbour Buildings.

(Countersigned) —————

Officer of Customs.

Dated 13th October 1896.

† A column headed "Final Destination of Goods" has been added since 1904.

E

used as packages, passengers' luggage, ships' stores, ballast, and military and naval stores on board Government vessels, goods transhipped *under bond,* and goods in transit through the country on a *through bill of lading* (of which separate accounts are given), and goods unlanded and so reported.

Under exports are included all goods entered on ships' bills of lading, excluding the classes after (*b*) in the previous paragraph; new ships, leaving our shores sold to foreigners are included since 1899.

Goods immediately reshipped at the same or another port, or held in bond and then reshipped, are included in imports, and in exports are distinguished as *Exports of Foreign and Colonial Produce.*

Bullion and coin are not included in the general totals of imports or of exports, but are recorded in separate tables. Coin carried privately and the great part of diamonds imported or exported (a quite important item) are not recorded.

The treatment of coal throws light on these paragraphs. Coal taken for use on the voyage is registered, but not included among exports; coal as cargo is included.

The value of imports reckoned is the nominal exchange value just before they are landed, and so includes all payments

Values.
due to foreigners, shippers, underwriters, etc., and shipping dues, and none to stevedores, dock-labourers, etc. The value of exports is the value "free on board." The exact definition of the values, here and in other countries, is of primary importance in studying the balance of trade.*

Great difficulty is experienced in classifying exports according to their countries of destination and imports according to their countries of origin; the details first asked for in 1904 (see notes on pp. 46 and 49) have led to greater accuracy and definiteness on these questions. In the accounts of trade there have been since 1904 two sets of tables, and the newer ones relating to countries of consignment are now given the greater importance.†

Very great care is necessary in using the accounts of foreign

---

* See the Reports of the Committee of the British Association on *The Accuracy . . . of . . . Statistics of International Trade,* 1904 and 1905.
† See *Committee on Trades Records* (Cd. 4346), and compare a current *Statistical Abstract of the United Kingdom* with those issued *circa* 1910 and 1903.

trade during the war period.   The class named above " military and naval stores on board Government vessels " excluded from the accounts assumed vast dimensions.

A very good example of an official inquiry is to be found in the Census of Production (1907) of which the results were published in 1912 (Cd. 6320).* Special attention may be directed to the relation between the *quæsitum*, the ultimate object of the inquiry, the *data* which it proved to be possible to collect, and the adjustment of the questions so that the answers could readily and accurately be given by the employers in various industries.

---

* Examples of the Blank Schedules used can be seen at the School of Economics.

CHAPTER IV.

TABULATION.

LEAVING now the consideration of blank forms of inquiry, let us turn to the methods by which our data, accumulated on these forms, can be tabulated. At first sight the tabulation of so many million census forms, so many schedules of wages, and so many lists of goods imported, seems mere office work, to be done mechanically, only requiring accuracy and not subject to scientific analysis. Tabulation does, indeed, involve a great deal of automatic labour; but the determination of the exact form of the table and the choice of the headings to which the totals shall correspond task the administrative statistician, and are worth the closest study.

The function of tabulation in the general scheme of a statistical investigation is sufficiently definite; it is to arrange in easily accessible form the answers to those questions with which the investigation is concerned. If it is required to know, for instance, the number of persons of each sex and age-group in all the districts of the country, the figures in the table must show these numbers. Or, to take a less definite problem, we want all the information possible as to annual earnings. In studying the forms issued for the Wage Census, we have seen that the information which can be obtained is not precisely that which we require. The problem then is so to tabulate our information that our totals may give answers as near to our requirements as possible, and it can easily be found by experiment that the way to do this is by no means obvious.

Not only must the figures be grouped so as to answer the questions put forward in the original scheme, but if the information is of wide and varied interest, as in all the inves-

tigations so far considered, the data must be studied from many points of view, and tabulated so that students in all branches of knowledge may be able to extract from our tables the information they require. Thus the population census is used by the financier, the legislator, the merchant, and the commercial traveller; political economists turn to it for light on the development of industry, and on the change of numbers in each trade; those interested in social questions will study the ages and sex-distribution in various districts or occupations; the sociologist and biologist will need accurate information as to the growth of population and the change of age distribution.

To take more specific points, the blue-book which contains the tabulation of foreign trade statistics will be expected to show how our trade with each country is developing, whether we are holding or improving our position in certain markets; whether we are exhausting our supply of raw materials; whether some new commodity is yet of importance. It must be remembered that the original material is not accessible to the public, that they are dependent on the information extracted for them, and that, though it would be possible to turn through all the forms for special data, yet the labour needed would be prohibitive, while a little more detail in the tabulation might easily have isolated the information needed.

The method of tabulation should be taken in relation to the conception of characteristics explained above (p. 20). Each person or thing in a group possesses certain Tabulation and adequately defined characteristics, say A, B, C, characteristics. and D. They also possess one or other of the characteristics E_1, E_2, E_3 . . ., and one or other of F_1, F_2, F_3 . . ., etc. A table in single tabulation shows separately the totals under each characteristic, E_1, E_2, etc. The heading of the table gives directly or by reference the definitions of A, B, C, and D, and contains frequently some such phrase as ·" in each locality " if the E characteristic is a locality. Each line in the first column then defines an E. / A double tabulation shows the classification both by E and by F, the heading of each column defining an F, so that an entry shows the number of persons who possess, say, the characteristics A, B, C, D, E_2, and F_6. The horizontal totals show the totals who have

characteristics E_1, E_2, etc., and the totals of the columns relate similarly to F_1, F_2, etc.

For convenience, the methods of tabulation may be divided into three groups : A. The simple statement of totals of ~Three groups of~ persons or things which satisfy given conditions, ~tabulations.~ such as the number living in a town, or the total value of imports from France; B. The grouping of a great number of units in relation to some particular property possessed by all, with the object, not of answering assigned questions, but of putting the material in a form ready for use in further investigations—*e.g.*, the population according to ages, or wage-earners according to the value of their wages; C. The tabulation of non-numerical answers in suitable groups to give a view of the whole—*e.g.*, the causes of strikes or the state of employment. The division between groups A and B is not always definite.

In the tabulation the convenience of the reader must be studied. The table must be so arranged that any totals required can instantly be found. This is to a great extent a question of typography, the use of suitable founts for figures and headings, and also of the choice of the right shape and size of page. Supposing the best possible choice made in these respects, our rule will then be to get the maximum amount of information into a given space.

GROUP A.—Thus we can have SINGLE tabulation, answering ~Classes of tabu-~ one or more groups of independent questions, ~lation.~ as :—

NUMBER AND MEMBERSHIP OF TRADE UNIONS.*

Year.	Number of Trade Unions at end of Year.	Total Membership of these Unions at end of Year.
1896	1,317	1,493,375
1897	1,307	1,611,384
1898	1,267	1,644,591

DOUBLE tabulation shows the subdivision of a total according to two categories, in the example giving on p. 55, according to sex and age :—

---

* Compiled from the *Sixth Annual Abstract of Labour Statistics*, p. 5.

CLASSIFICATION OF PAUPERS IN IRELAND.—Total Numbers who received Relief during the Year ended Lady Day 1892.*

Ages of Persons Relieved.	Males.	Females.	Total.
Under 16 years · · ·	44,391	43,648	88,089
Of 16 and under 65 years ·	132,370	79,045	211,415
Of 65 years and upwards ·	35,121	45,668	80,789
All ages · · ·	211,882	168,861	380,243

More information may be included thus :—

CLASSIFICATION OF PAUPERS IN ENGLAND AND WALES.—Total Numbers who received Relief during the Year ended Lady Day 1892.†

Ages of Persons Relieved.	Indoor.	Outdoor.	Total.	Metro-polis.	Other Parts of England and Wales.
Under 16 years · ·	111,782	441,805	553,587	100,671	452,916
Of 16 and under 65 years	232,284	385,299	617,583	148,066	469,517
Of 65 years and upwards	114,144	287,760	410,904	64,779	337,125
All ages · ·	458,210	1,114,864	1,573,074	313,516	1,259,558

A TREBLE tabulation can be used, subdividing the total into three distinct categories, with cross totals for each group. Thus the table on p. 56 gives separate divisions according to age, sex, and district; percentage lines, in a distinct type, are also introduced :—

The same process can be further extended : the example in the table opposite shows an arrangement for a QUADRUPLE tabulation, distribution by district, date, sex, and industry, with subsidiary information ; but it is generally better to use two or more tables than to increase the complication, unless it is necessary to bring several categories into close relation. Suitable varieties of type will often make comparisons easy in a very complex table.

---

* Compiled from the *Sixth Annual Abstract of Labour Statistics*, p. 102.
† *Ibid.*, p. 101.

CLASSIFICATION OF PAUPERS BY AGE, SEX, AND LOCALITY.*

Total Numbers who received Relief in England and Wales during the Year ended Lady Day 1892.

Ages of Persons Relieved.	METROPOLIS.				OTHER PARTS OF ENGLAND AND WALES.				TOTALS. ENGLAND AND WALES.			
	Males.	Females.	Total.	Percentage at each age.	Males.	Females.	Total.	Percentage at each age.	Males.	Females.	Total.	Percentage at each age.
Under 16 years†	...	...	100,671	32.1	...	...	462,916	35.9	...	...	553,587	35.2
Of 16 and under 65 years	74,207	73,859	148,066	47.2	202,180	267,337	469,517	37.3	276,387	341,196	617,583	39.3
Of 65 years and upwards—												
Number	27,238	37,541	64,779	20.7	136,392	200,733	337,125	26.8	163,680	238,274	401,904	25.5
Percentage of sexes	42.0	58.0	100	...	40.4	59.6	100	...	40.7	59.3	100	...
All Ages	...	...	313,516	100	...	...	1,259,558	100	...	...	1,573,074	100

Ibid., 101.   † The returns do not distinguish sex under 16 years.

Looking now at the census householders' schedule (facing p. 22), we can see that there are about thirteen different items of information about each person : district, posi- *Tabulation of tion in family, condition as to marriage, children, census material. sex, age, occupation, industry, industrial status, infirmity, birthplace, nationality, and house-room. These could be tabulated in 78 different double, 286 treble, or 715 quadruple tabulations, so that there is plenty of scope for choice.

To fix our ideas, we will take occupation as the main sub-division, and examine Mr. Booth's use of the Mr. Booth's census returns, say for London Printers.* tabulation.

First he gives a treble classification—occupation, sex, and age—using columns corresponding to 3, 4 and 10 of the 1911 schedule.

Census Divisions, 1891	FEMALES.	MALES.			TOTAL.
	All Ages.	·19.	20-54.	55·	
1. Printer - · ·	1,316	9,988	21,784	1,921	35,009
2. Lithographer, &c. -	809	757	3,037	437	5,040
Total - ·	2,125	10,745	24,821	2,358	40,049

Then follows a single table, district and numbers, using the information on the back of the schedule.

DISTRIBUTION.

E.	N.	W. & C.	S.	TOTAL.
5,884	9,835	7,577	16,753	40,049

Three simple tables are then given, relating to heads of families, using columns 2, 3 and 4 (sex), 2 and 14 (birthplace), and 2 and 12 (industrial status).

---

* *Life and Labour of the People*, vol. vi., p. 189.

His next table uses columns 2 and 10, and is as follows :-

	Heads of Families.	Others Occupied.	Unoccupied.	Servants.	TOTAL.
			TOTAL POPULATION CONCERNED.		
Total - - -	18,048	16,060	47,257	854	82,219
Average in Family -	1	.89	2.62	.05	4·56

The next table (not here given) is a single classification according to number of rooms and servants, a most ingenious indirect use of the scheduled information; and the last is an example of the legitimate use of a quadruple tabulation—occupation, industrial status, sex, and age—given on the next page.

It would be difficult to find a better example of tabulation of a great multitude of details to serve a special purpose. The The census census authorities had in many cases not tabu-tabulations. lated the necessary details, and it was necessary to turn through the original schedules to get at the facts. For such work as this, the function of tabulation is simply to provide the answers to definite questions. Thus the census reports show how many persons of each sex and age-group belong to certain industries in certain places, in a quadruple tabulation extending over many pages, each page relating to one district, and this table may be used for accomplishing many separate purposes : each item is already a total ready for use. It is impracticable from limits of time and space, even if it were desirable, to tabulate all the possible groups of qualities which can be made from all the statements on each census form; a good tabulation will aim at providing only those statements which are of practical use. Thus many simply descriptive totals are given, such as the numbers of each sex and age in each parish in the United Kingdom, to serve primarily for administrative purposes; and many statements which will afford the economist and sociologist the opportunity of tracing the progress of industries, of studying the ages of workpeople in different occupations, the changes in age-grouping of the nation; and some further tables might

STATUS AS TO EMPLOYMENT (according to Census Enumeration).

Census Divisions (1890)	Employers		Employed				Neither Employer nor Employed		Total
	Males.	Females.	Males.		Over 20	Females of all Ages.	Males.	Females	
			Under 20	Over 20					
1. Printer	827	39	9,988	22,565		1,266	313	11	35,009
2. { Litho., Copper, and Steel-plate Printer	177	2	506	2,571		88	153	6	3,503
{ Map and Print Colourer and Seller	17	...	49	175		72	62	12	387
{ Ticket and Label Writers	36	3	202	169		619	114	7	1,150
Total	1,057	44	10,745	25,480		2,045	642	36	40,049
	1,101		38,270				678		

Proportion of Employers to Employed— 1 to 35.

no doubt be given to throw light on problems of special interest. In each successive census new tables are to be found.

It is interesting to open one of these great tables of figures, such as are generally to be found forming the bulk of a blue-book, and taking a figure at random, ask " Why is this figure printed, what question does it answer, to whom can it give information ? " For instance, in the *Eighth Report on Trade Unions*, p. 257, we find that the United Brickworkers' and Brick Wharf Labourers' Union spent £20 on funeral expenses in 1894, an average of 3s. 7½d. per member. As an isolated statement this may interest a very small number of persons; but that small number has a right to expect that they shall find the figures relating to their union tabulated in a general official book; to them it may be as important as the item, on the same page, of £5,481 spent by the Boiler-makers. From this point of view, the question of inclusion of such small items is simply one of space. If space is limited, a selection would be made of larger quantities only, as being likely to concern more people.

Minutiæ.

But there is a reason of quite another character for printing such items as these. The raw material, on which the totals in such tables are based, is not accessible to the student except by means of this Report. Now, the compiler of these statistics cannot know from what particular point of view they will be studied. It may be desired to examine and group trade unions according to their expenditure on different items, to study their history, classifying them as fighting organisms and as friendly societies. The tabulations needed cannot well be foretold. The material is therefore given in the rough, in order that the tabulation may be made by each student according to his needs. At the same time the most suggestive totals are given as one of these possible methods of tabulation; and in the summary of such a report, the items are retabulated, the rough material being omitted, in those ways which the editor thinks most useful.

Importance of raw material.

When space is much too limited for any publication *in extenso* of the items, a careful selection must be made of those to be printed; and it is this selection that is generally open to most criticism.

Selection of raw material.

The Census supplies an illustration from the County Borough

of Coventry, 1911,* where the following detail is given for 115 persons :—

BRICK, CEMENT, POTTERY AND GLASS.   MALES.

Age · ·	10-	13.	14.	15.	16.	17.	18.	19.	20-	25-	35-	45-	55-	65-	Totals.
Workers · · ·	—	—	2	3	2	3	1	4	12	36	23	17	4	—	107
Dealers · · ·	—	—	—	—	—	—	—	—	1	4	—	1	1	1	8

while all the males—masters, foremen, skilled workmen, labourers and boys—engaged in the cycle and motor-car trade are shown in no more detail than :—

VEHICLES.

Age · ·	10-	13.	14.	15.	16.	17.	18.	19.	20-	25-	35-	45-	55-	65-	Totals.
Cycle and Motor Car— Makers, Mechanics	—	1	192	271	303	325	371	372	2,003	3,872	2,488	1,122	379	76	11,775
Motor Car— Makers, Mechanics	—	—	*70*	*108*	*143*	*160*	*210*	*208*	*1,299*	*2,524*	*1,376*	*537*	*158*	*36*	*6,898*
Others · · ·	—	—	—	2	1	4	8	7	31	62	49	24	21	4	213

It is explained on p. iii of the volume that full particulars (by age) of relatively important occupations in a district are shown in italics.

In such cases, two useful rules might be applied : omit all numbers under, say, 500 when by so doing a *line* of print would be saved ; and give all numbers over 10,000 correctly only to the nearest 100, and so for other digits in proportion, thereby reducing the width of *columns* of print.   If, for example, we knew to the nearest 100 the exact numbers in each district and occupation in which as many as 1000 were employed, our knowledge would be as complete   Economy of space. as we needed ; and it is doubtful whether the space occupied by such a tabulation would be more than that already devoted to the subject.   In many cases, on the other hand, it is essential to have the raw material quite unchanged.   Each tabulation must be judged on its own merits.

If may be useful to take a particular group of answers, and discuss what tabulations will throw most light on the questions at issue.   The Poor Law Commissioners of 1833   Tabulation of the Poor Law Returns, 1833. collected information from a thousand villages in England and Wales on the following six points

* Census Report, Cd. 7019, p. 597.

among others : the wages of an agricultural labourer in summer and in winter, both with and without the inclusion of beer as part payment, his annual earnings, and the subsidiary earnings of his wife and children. It may be supposed that the chief object of the Commissioners was to find whether the labourers' families earned enough for their support, and what proportion was earned by the wives and children.

The following scheme of tabulation would show in what *counties* the labourer was badly off :—

| County. | Average Annual Earnings of | | |
	Man.	Family.	Together.

The counties might be taken in alphabetical order for convenience of reference, or in geographical order with subordinate averages for groups (*e.g.*, Eastern : Norfolk, Suffolk, Essex); or the counties might be arranged in the order of the total earnings, so that it could be seen at a glance in which counties the labourers were worst off.

To show the number of villages, county by county, in which the earnings were below a certain minimum, or within certain limits, the table given on p. 63 might be used.

This table can be used in the above complex form or simplified. The number of subdivisions of money to be distinguished depends on the space at disposal and on the number of villages which would be entered in each. A table in which most of the entries are 1 or 0 is open to criticism. In the above table the villages are too few to allow accuracy in percentage.

It will be seen that this table would furnish the answer to almost all questions which could be put as to total earnings. Tabulation to show correlation. For instance, if we wish to see the relation between total earnings and the family's subsidiary contribution, we should look at the smallest totals in the last column but one and see if they corresponded with the largest percentage of family earnings. If we found signs of corre-

spondence, we should rearrange the counties in the order of these subsidiary percentages, and see if they were approximately in order of total earnings also. This is an example of tabulation to show correlation, the correspondence in the occurrence of two sets of phenomena.

Another important group of questions arising in connection with these tables is : What is the relation between weekly wages and annual earnings, and what proportion   Wages and of the wage is generally paid in kind? We shall   earnings.

ANNUAL EARNINGS OF MEN AND FAMILIES.

		Number of Villages in which the Total Earnings averaged						Average Earnings in County of			Family Earnings as Percentage of Total
	£25.	Above £25 and not above £30.	Above £30 and not above £35.	Above £35 and not above £40.	Above £40 and not above £45.	Above £45 and not above £50.	Above £50	Man.	Family.	Total.	
IN NORFOLK -	0	1	3	6	4	3	2	£30	£11	£41	27
Percentages of Total Number of Villages -	0	5	16	31½	21	16	10½	...	...	...	...
IN SUFFOLK -	0	3	4	5	3	2	2	£28	£11	£39	28
Percentages of Total Number of Villages -	0	16	21	26	16	10½	10½	...	...	...	...
IN ESSEX - -	1	3	6	7	10	3	1	£28	£10	£38	26
Percentages of Total Number of Villages -	3	10	19	23	32	10	3	...	...	...	...
In Eastern Counties	1	7	13	18	17	8	5	£28 10	£10 10	£39	27
Percentages of Total Number of Villages -	1	10	19	26	25	12	7	...	...	...	...

not now require the statements as to subsidiary family earnings. In records of agricultural wages the most common statement was, *e.g.*, " wages in this district are from 10s. to 12s. a week." Now, a farm labourer did not generally earn as much in winter as in summer, because wages were reduced to correspond to the smaller amount of work necessitated by failing light ; from this cause annual earnings will be less than the weekly wage multiplied by 52. Besides this wage he generally receives special money at hay and wheat harvests, and also many

payments in kind, such as daily beer, house and ground at reduced rent, and other privileges. It is generally best to value all these, and compute his earnings thus :—

10s. for 38 weeks -	-	£19	0	0
12s. for 9 weeks (summer)		5	8	0
Hay harvest, 1 week	-	0	15	0
Wheat harvest, 4 weeks		5	0	0
Beer, 1s. per week	-	2	12	0
Cottage and ground	-	5	0	0
Other perquisites -	-	1	5	0

£39 0 0 = 15s. per week.

In this case earnings are 50 per cent. above the general weekly wage. An estimate of this nature has been made by the late Mr. Little for each county for 1867–70 and 1892.

The question, Are winter wages generally below summer wages, and by how much? can be answered by the following scheme of tabulation, which uses the data not employed in the previous tables :—

Winter and summer wages.

COUNTIES.	Average Weekly Wage in		Number of Villages where the Excess of Summer Wages over Winter was					
	Summer.	Winter.	Nothing.	6d.	1s.	1s. 6d.	2s.	More than 2s.
	s. d.	s. d.						
Norfolk · · ·	11 2	10 3	13	2	3	2	5	3
Percentage of Number of Villages included · · · · ·			*46*	*7*	*11*	*7*	*18*	*11*
Suffolk · · ·	10 2	9 8	24	0	6	1	26	1
Percentage of Number of Villages included · · · · ·			*70*	*0*	*18*	*3*	*6*	*3*
Essex · · · ·	10 9	9 10	22	0	11	0	5	4
Percentage of Number of Villages included · · · · ·			*52*	*0*	*26*	*0*	*12*	*10*
EASTERN COUNTIES	10 6	9 11	59	2	20	3	12	8
Percentage of Number of Villages included · · · · ·			*57*	*2*	*19*	*3*	*12*	*8*

These examples do not quite exhaust the useful tabulations of these groups of figures, for we have not yet examined the distribution of wages, that is the relative numbers paid at different rates. These returns do not, however, illustrate such a tabulation well, for we are not told the rates paid to individuals, but only the rate prevalent in the villages.

GROUP B.—The grouping according to wages affords an example of the second method of tabulation. We have now no definite questions to answer, as in the method so far discussed, but a more general problem : given a mass of data, it is required to tabulate it, so as to present the maximum amount of useful information. Our raw material is so many thousand isolated statements, which must be focussed, made to present definite meaning, and worked up so as to be useful for future comparison.

Some investigations are undertaken not to answer any definite questions or to throw light on any given problem, but to collect information which, though it has no immediate use, is likely to be needed ulti- mately by many investigators occupied with various questions. Such is a wage census. So long as we have no sufficient account of wages, we are badly informed as to one of the most important measurements of the social body, and economists and statisticians are continually hindered by the want of data essential for their work; but the census has no immediate practical use, for knowing the height of wages does not help us directly to regulate that height. In such an investigation our object will be to examine the figures, and give all the groupings and averages which seem likely to be useful for any purpose; and while doing this we shall imperceptibly pass to a different class of investigation; we shall be finding a structure underlying our multifarious details; we shall find that the chaos, which our figures present at first sight, obeys laws; we shall be making a visible outline, and giving a definite shape to our apparently featureless mass.

The complete discussion of this problem belongs to a later chapter; but the tabulation can be begun without special technique. The examples taken will relate chiefly to wages, but the methods are quite general.

In the American *Report on Wholesale Prices, Wages and Transportation* of 1891, the wages of some 10,000 persons are

F*

detailed? It is proposed to consider their tabulation as a
Selection of limits homogeneous group. The results are given on
of groups. pp. 69, 70. In the original publication the wages
are given to half a cent; in the second column, on p. 69, the
numbers of wage-earners are given in 10-cent groups, from
$.25 to $.34, $.35 to $.44, and so on, those earning wages
exactly at the dividing points being always placed in the
division below. Notice that the average wage of such a group
as $2.15 to $2.24 is not $2.20 if the wage-earners are evenly
distributed cent by cent, but the average of $2.15, $2.16, . . .
$2.24, *i.e.*, $2.195.

Looking at column 2, we shall see that the figures present
no order, follow no rule; no structure has yet been found, our
divisions are too narrow for our material.

Now group the wage-earners with wider limits, as in
column 6, where the numbers earning in half-dollar groups
are given; we have here a nearly regular sequence of numbers
falling after the maximum in the second group. Going back
to narrower limits, to find exactly at what divisions this regu-
larity is first in evidence, we have in column 4 the numbers in
20-cent groups which show considerable, but not absolute
regularity. The numbers in 30-cent groups * are successively
75, 355, 674, 1242, 740, 660, 343, 310, 180, 181, 233, 32, 82,
3, 4, 8, 1, almost completely regular except for the large group
at $3.25 to $3.55.

The question as to which of these groupings should be
selected is to be decided by the number of separate items the
eye can instantaneously grasp. In looking at the 51 numbers
in the 10-cent groups, or the 26 in the 20-cent, the meaning is
lost in a maze of figures (though as many details as these
could be properly shown in a diagram), but the 11 numbers
in the half-dollar groups are easily comprehended.

Stated in words, the result of our tabulation (column 7) is
that 6 per cent. of the wage-earners made from $.25 to $.74,
29 per cent. from $.75 to $1.24, and so on.

For the practical work of the tabulation from the original
figures, we should take ruled sheets, enter at the head of
Practical tabu- successive columns certain wage limits, and
lation. turning through the items enter each wage by a

---

* *Vide*, p. 97, *infra*.

dash in its appropriate column, grouping them in fives and tens, to facilitate addition.

From the preceding paragraphs it is clear that we do not need to take separate columns for each cent from $.25 to $5.35 for tabulation, but a little consideration is necessary to see how minute the limits should be to give the correct average.

Suppose the entries in cent groups to be :—

$1.70	$1.71	$1.72	$1.73	$1.74
IIIII I	IIIII IIIII IIIII II	IIIII IIIII III	IIIII IIIII	IIIII

The average of the wages so entered can be quickly calculated as $1.718.

If, on the other hand, we put all the 51 entries as simply " between $1.70 and $1.74," or more exactly " as much as $1.70 but less than $1.75," we should naturally take them to be all (for purposes of averaging) at the middle point of this group, viz., $1.72.

If we have a sufficient number of items, the differences between the average assumed and that calculated for each group will be very slight. This is seen on p. 69; column 8 gives the averages calculated from the entries in 10-cent groups, while column 9 gives them on the hypothesis that for purposes of averaging the numbers in the half-dollar groups may all be taken at the middle points of their groups. The difference is greatest in the first and last, the smallest groups. The general average obtained from column 9 is $1.70, which is the nearest round number to the true average $1.73. Hence, for the purpose of obtaining the general grouping and average, we need only take 11 half-dollar columns for marking in our items.

For other purposes it may be advisable to work more minutely; for in the lowest group, we shall wish to know how many are earning $.25, $.30, $.35 separately, for 5 cents is a perceptible difference on 25 cents. At the top also it may be useful to know the exact wages.

More minute entries again will be needed for the second method of tabulation, which is as follows :—Suppose all the

wage-earners to be arranged in order of the magnitude
The Galtonic of their wages, those at $.25 at one end, those
method. at $5.75 at the other. Note the wages of men at
given points in the row. The lowest wage is $.25; one-tenth
of the way along, that of the 512th worker is between $.85 and
$.95, . . . half-way up the wage is $1.50. The figures at each
tenth are given on p. 70. By this means we get a very vivid
idea of the distribution according to wages.

These numbers cannot be obtained accurately if we have
only entered the details correct to half-dollars, but can be
found from the 10-cent grouping, which is therefore the classi-
fication to be adopted. We must first determine in which of
the small groups the men one-tenth, two-tenths . . . up the
group lie, and then estimate their position inside the smaller
group. Thus, if we want the figure more accurately than
" between $.85 and $.95," as given above, we proceed as
follows :—The 512th man from the bottom is the 82nd man in
the group between $.85 and $.95, for there are 430 earning less
than $.85; this group contains 169; if they were distributed
regularly, 17 to each cent, the 82nd man would be half-way
through this group, between $.89 and $.90. The hypothesis
of even distribution is sufficiently correct for most purposes,
and this method affords a sufficiently accurate means of
determining the wage of the workers at the tenth places. The
resulting figures are given on p. 70. If, however, we want to
know the wage of the half-way man more exactly, we see from
the half-dollar groups that it is between $1.25 and $1.75, a
rough approximation shows it to lie probably between $1.45
and $1.55, and then we rapidly turn through our original data,
isolating the wages at $1.46, $1.47, . . . $1.55.*

A slight modification of this method is also useful. Take
the average of the lowest 512 (or tenth), namely, $.70¼; of the
next, namely, $1.03; and so on (see p. 70). These figures also
give a vivid view, and are very convenient for comparisons
with other groups.

The figures so far apply to only half of the data in the
Senate Report. On p. 70 the whole are tabulated to give
the average wages of the successive tenths. A comparison of
the two groups so obtained shows how far the first half was
typical of the whole.

TABULATION OF WAGES—AMERICAN FIGURES, 1891.

1. Earning Daily Wages. ($) as much as / and less than	2. No. of Persons.	3. ($) as much as / and less than	4. No. of Persons.	5. ($) as much as / and less than	6. No. of Persons.	7. Percentage.	8. Average Wage in Group.	9.
.25 .35	1							
.35 .45	15	.25 .45	16				$ instead $	
.45 .55	59						.62 of .50	
.55 .65	85	.45 .65	144	.25 .75	317	6.2		
.65 .75	157							
.75 .85	113	.65 .85	270					
.85 .95	169							
.95 1.05	201	.85 1.05	370	.75 1.25	1,472	28.7	1.09	1.00
1.05 1.15	304							
1.15 1.25	685	1.05 1.25	989					
1.25 1.35	99							
1.35 1.45	458	1.25 1.45	557					
1.45 1.55	466			1.25 1.75	1,297	25.3	1.49	1.50
1.55 1.65	72	1.45 1.65	538					
1.65 1.75	202							
1.75 1.85	329	1.65 1.85	531					
1.85 1.95	58							
1.95 2.05	273	1.85 2.05	331	1.75 2.25	970	18.9	1.99	2.00
2.05 2.15	45							
2.15 2.25	265	2.05 2.25	310					
2.25 2.35	33							
2.35 2.45	101	2.25 2.45	134					
2.45 2.55	196			2.25 2.75	506	9.9	2.53	2.50
2.55 2.65	13	2.45 2.65	209					
2.65 2.75	163							
2.75 2.85	2	2.65 2.85	165					
2.85 2.95	15							
2.95 3.05	129	2.85 3.05	144	2.75 3.25	198	3.9	3.04	3.00
3.05 3.15	5							
3.15 3.25	47	3.05 3.25	52					
3.25 3.35	12							
3.35 3.45	0	3.25 3.45	12					
3.45 3.55	221			3.25 3.75	254	5.0	3.51	3.50
3.55 3.65	5	3.45 3.65	226					
3.65 3.75	16							
3.75 3.85	11	3.65 3.85	27					
3.85 3.95	0							
3.95 4.05	82	3.85 4.05	82	3.75 4.25	96	1.9	4.00	4.00
4.05 4.15	0							
4.15 4.25	3	4.05 4.25	3					
4.25 4.35	0							
4.35 4.45	0	4.25 4.45	0					
4.45 4.55	3			4.25 4.75	4	0	4.50	4.50
4.55 4.65	1	4.45 4.65	4					
4.65 4.75	0							
4.75 4.85	0	4.65 4.85	0					
4.85 4.95	0							
4.95 5.05	8	4.85 5.05	8	4.75 5.25	8	.2	5.00	5.00
5.05 5.15	0							
5.15 5.25	0	5.05 5.25	0				5.35	5.25
5.25 5.35	1	5.25 5.35	1	At 5.35	1			
Totals	5,123		5,123		5,123	100		
Average Wage	$1.731						Average Wage $1.70	

Wages of " Tenth " Men (*deciles*).		
Lowest Wage	.	$.25
1⁄10th up Group	.	.89
2⁄10th ,,	.	1.12
3⁄10th ,,	.	1.22
4⁄10th ,,	.	1.39
5⁄10th ,,	.	1.49
6⁄10th ,,	.	1.75
7⁄10th ,,	.	1.99
8⁄10th ,,	.	2.36
9⁄10th ,,	.	2.98
Highest Wage	.	5.35

(Wages of Men)

Average Wage of		Same for 10,000 Workers.
Lowest tenth -	$.70	.79
Second ,, -	1.03	1.00
Third ,, -	1.18	1.24
Fourth ,, -	1.28	1.50
Fifth ,, -	1.44	1.50
Sixth ,, -	1.59	1.88
Seventh ,, -	1.86	2.00
Eighth ,, -	2.14	2.22
Ninth ,, -	2.59	2.58
Highest ,, -	3.51	3.55
General Average 1.731		1.82

The tabulation of the data collected for the WAGE CENSUS of 1886 on such forms as that on p. 71, illustrates well some of the difficulties involved. The items given on the main part of the schedule are of this kind :—

No. Average Wage.
Spinners—Time : 6 : 12s. : 56½ hours.

Such returns are not perfectly definite, for if many are employed in the same occupation in a mill, it is possible that **Tabulation in** they will earn at different rates. Thus this entry **the wage census.** of 6 at 12s. might arise from either 6 men each earning 12s., or 2 at 10s., 2 at 12s., 2 at 14s. (average 12s.); or 4 at 12s., 1 at 15s., 1 at 11s. ; or 5 at 12s. and 1 at 18s.—12s. being the general rate, but not the average, in these last two alternatives. Since the purpose of the wage census was to give a comprehensive account of wages adapted for use in all investigations, it should show the numbers in all trades and subdivisions of employment by age, sex, and district, the average and general rate of pay for each group, and sufficient details to show the distribution about the average in each group, for a mere average may conceal exceptionally high or exceptionally low wages.

On inquiry at the Labour Department as to whether the original information had been given in a more detailed form than the line above, or whether divergencies might be concealed, the author learnt that the subdivision of occupations had been carried to such an extent, that in practice, where there was any great variation in the wages of workers under one heading, that heading had been split up, so that each

RATES OF WAGES.*

SILK MANUFACTURE.

Average Rates of Wages of Operatives Employed in the Silk Manufacture in Various Districts, at 1st October 1886.

DISTRICT I.—CHESHIRE, STAFFS, NOTTS, DERBY, AND WARWICK.

Total Number of Persons employed at 1st October 1886, as per Returns, 5,279.

Total amount paid in Wages in 1885, as per Returns, £129,588.

OCCUPATIONS.	Number employed as per Returns.	Percentage that the Number returned under each Occupation is of Total Number of Persons returned.	AVERAGE Weekly Rate of Wages.	Number returned at Rates within 10 per cent. above and below the average.	Average *Maximum* Weekly Rate of Wages.	Number on which Average Maximum Rate is based.	Average *Minimum* Weekly Rate of Wages.	Number on which Average Minimum Rate is based.
			Total Number, 1,085.	MEN.		Hours of Labour per week, 56½.		
			s. d.		*s. d.*		*s. d.*	
THROWN SILK—								
Overlookers and Stewards · · · Time	79	1.5	23 3	35	31 3	4	20 0	29
Assistant Overlookers · · · Time	9	.2	17 3	6	19 0	...	15 0	...
Spinners · · · Time	24	.5	11 9	15	13 0	5	9 3	4
Millmen (Throwsters) · · {Time	109	2.1	17 7	68	20 11	18	14 0	3
{Piece	9	.2	18 8	7	...	...	...	...

* From *Wages in the Minor Textile Trades* (C.—6161 of 1890).

group was separately entered, or that several groups were distinguished under one heading; and that when there was reason to believe from the light of other returns that this had not been done, supplementary inquiries were made on this point, so that the original data were detailed enough for any requisite fineness of tabulation.

The problem then was to tabulate the answers from the various factories in a district, to show clearly and succinctly the distribution of wages in each subdivision and in the whole. It can hardly be said with confidence that the method adopted, of which a specimen is given on p. 71, is entirely satisfactory.

To clear our ideas let us suppose that the details on which the line relating to throwsters (time) was based were as follows :—

3 earning	14/	-	" average minimum rate."
14 ,,	15/		
6 ,,	15/6		
20 ,,	16/	⎫	
10 ,,	17/6	⎪	68 within 10 per cent. of the average
20 ,,	18/	⎬	for all, which is 17/7.
8 ,,	18/6	⎪	
10 ,,	19/	⎭	
10 ,,	20/6	⎫	18 earning 20/11 on the average.
8 ,,	21/5	⎭	

The process adopted in the tabulation may be supposed to have been to separate from the whole group of returns a small group of old men or inferior workers earning far below the average, and enter them as a distinct minimum group, and to separate a small group of the most skilled workers and enter them as a maximum group. This is better than giving simply the highest and lowest of the individual wages, for either of these may be due to exceptional circumstances, and may be quite a long way from that paid to any other person. The exact size of these extreme· groups must be determined from inspection of the returns themselves. After this has been done, the remaining wages may not be grouped close together; in the example taken they are scattered between 15s. and 19s. To give some clue as to this distribution the number earning within 10 per cent.

Various methods possible.

of the average is stated; this is probably the best way if only one column can be devoted to it, but 10 per cent. is a wide limit to adopt. Another method would be to give the limits within which the wages of the 10 per cent. of the earners above and 10 per cent. below the average were contained : in this case 16s. and 18s.

If, however, not more than 8 columns are to be devoted to each group, the following arrangement would give much more definite information, and it could have been made from the data in hand, and would be well adapted for all the purposes for which it would be required.

Number employed - - - - - - 109
Average weekly rate - - - - - - 17/7
One-tenth of the number of wage-earners
    received not more than - - - - 15/
One-quarter of the number of wage-earners
    received not more than - - - - 16/
One-half of the number of wage-earners
    received not more than - - - - 18/
One-quarter received not less than - - - 19/
One-tenth   ,,   ,, ,, ,, - - - 20/6

This method was used in the publications of the wage-census of 1906, except that the tenths were not given.

After studying Chapters V and VI, readers will naturally replace the phrases used above by the terms median, quartiles and deciles, and consider whether one of the measures of dispersion would not be more appropriate to use than the details here suggested.

We are fortunately not dependent solely on the tabulation as given above, for wages in industries as a whole are also tabulated for 1886 * on the following plan, which is in a form most useful for purposes of comparison (p. 74).

  The lines giving percentages are very helpful. We can at a glance compare the levels of wages in different industries. Thus in the cotton manufacture the average wage is 2s. higher than in the woollen; and in the cotton there is a large group of highly skilled workers earning from 30s. to 35s., while in the

The general summary.

---

* • More detail is shown in the Reports for 1906.

NUMBER AND PERCENTAGE OF PERSONS EMPLOYED AT VARIOUS RATES OF WAGES.*

Table showing the average Normal Wages paid to men in the undermentioned employments, and the Number and Proportion of men paid at different rates, at October 1886.

		Under 10s.	Of 10s. and under 15s.	Of 15s. and under 20s.	Of 20s. and under 25s.	Of 25s. and under 30s.	Of 30s. and under 35s.	Of 35s. and under 40s.	Above 40s.	TOTAL.	Average Wages per Head. s. d.
Cotton Manufacture	Number	2	370	8,793	8,822	4,525	7,283	1,582	812	32,189	25 3
	Per cent.		1.2	27.3	27.4	14.1	22.6	4.9	2.5	100	
Woollen „	Number	...	146	3,377	5,559	1,725	705	392	344	12,248	23 2
	Per cent.	...	1.2	27.6	45.4	14.1	5.7	3.2	2.8	100	
Worsted and Stuff Manufacture	Number	...	835	1,705	909	2,635	879	28	14	7,005	23 4
	Per cent.	...	11.9	24.3	13.0	37.6	12.6	0.4	0.2	100	
Linen Manufacture	Number	192	780	2,952	2,070	416	290	39	68	6,807	19 9
	Per cent.	2.8	11.4	43.4	30.4	6.1	4.3	0.6	1.0	100	
Jute „	Number	...	565	1,038	964	127	53	52	...	2,799	19 4
	Per cent.	...	20.2	37.1	34.4	4.5	1.9	1.9	...	100	
Hemp „	Number	...	25	300	581	168	39	94	25	1,232	23 6
	Per cent.	...	2.0	24.4	47.2	13.6	3.2	7.6	2.0	100	
Silk „	Number	...	324	881	367	278	121	273	4	2,248	22 3
	Per cent.	...	14.4	39.2	16.3	12.4	5.4	12.1	0.2	100	
Carpet „	Number	...	...	130	183	834	100	15	30	1,292	26 7
	Per cent.	...	...	10.1	14.2	64.5	7.7	1.2	2.3	100	
Hosiery „	Number	...	...	296	458	51	190	75	...	1,070	24 5
	Per cent.	...	...	27.7	42.8	4.8	17.7	7.0	...	100	

* General Report on Wages (C.—6889 of 1893).

woollen nearly half are close to the. average, earning' between
20s. and 25s. In the jute and linen manufactures the averages
are nearly the same, but in the former a larger proportion are
below the 15s. limit. In the silk manufacture there is an
aristocracy as in the cotton, but it is smaller and better paid,
for 12 per cent. earn more than 35s. This table is a master-
piece of concentration and clearness.

. We will discuss next the tabulation of the figures relating
to CHANGES in RATES of WAGES collected by the Labour De-
partments. The following examples are taken   Tabulation of
from the earliest report; the form of the tables   change of
has been modified many times since then, and a   wages returns.
study of these alterations can be usefully followed by turning
through a file of the annual reports. The details collected
on the earlier blank forms show the occupations and numbers
affected, the dates from which the changes took place, and
the wages and hours in a full week exclusive of overtime (a
definition corresponding exactly to that used for the wage
census) before and after the change.

EXTRACT FROM TABLE showing the Changes in Rates of Wages and
   Hours of Labour of Ordinary Agricultural Labourers in Various
   Districts of the United Kingdom in 1894, so far as reported to
   the Board of Trade.*

County and Union.	Particulars of Changes in Summer Wages. (1894 compared with 1893.)		Particulars of Changes in Winter Wages. (1894 compared with 1893.)		No. of Male Agricultural Labourers, Farm Servants, Shepherds, Horsekeepers, Horsemen, Teamsters, Carters, in '91.
	Increase.	Decrease.	Increase.	Decrease.	
		Per Week.	Per Week.	Per Week.	
LINCOLNSHIRE—					
Gainsborough -	...	...	...	1/6(15/to 13/6)	2,466
Louth - -	...	...	...	1/6(13/6to 12/)	3,932
Spilsby - -	...	...	...	1/6(13/6to12/)	3,288
NORFOLK—					
Aylsham - -	...	1/(12/ to 11/)	...	—	2,576
Docking - -	...	6d.(12/6to 12/)	1/(10/-11/)	...	2,487
Flegg, East and West -	...	1/(12/ to 11/)	...	1/(11/to 10/)	1,108
Forehoe - -	...	...	...	1/(11/to 10/)	1,448

* From the second *Annual Report on Changes of Wages*, pp. 198–9; a little
compressed.

EXTRACTS FROM TABLE showing the Changes in Rates of Wages of Ordinary Agricultural Labourers in Various Districts of the United Kingdom in the Summer of 1895, so far as reported to the Board of Trade.*

County and Union.	No. of Male Agricultural Labourers, Farm Servants, Shepherds, Horsekeepers, Horsemen, Teamsters, Carters, in 1891.	Particulars of Changes in Summer Wages (1895 compared with 1894). *Decreases in italics.*	Weekly Rate of Wages in Summer.			
			1894.		1895.	
		Per Week.	*s.*	*d.*	*s.*	*d.*
DURHAM—						
Stockton* · ·	437	*Decrease of 6d.*	17	6	17	0
Teesdale · ·	669†	Advance of 6d.	17	6	18	0
(Barnard Castle Rural Dist.).*						
OXFORDSHIRE—						
Headington · ·	1,118	*Decrease of 1s.*	12	0	11	0
Henley · ·	1,587†	*Decrease of 1s.*	12	0 to	11	0 to
(Hambleden Rural Dist., Bucks).			14	0	13	0
NORFOLK—						
Flegg, East & West	1,108	*Decrease of 1s.*	11	0	10	0
Forehoe · ·	1,448	*Decrease of 1s.*	11	0	10	0
Henstead · ·	1,504	*Decrease of 1s.*	11	0	10	0
Mitford and Launditch · ·	3,622	*Decrease of 1s.*	11	0	10	0
Smallburgh · ·	2,264‡	*Decrease of 1s.*	11	0	10	0
Swaffham · ·	1,942	*Decrease of 1s.*	11	0	10	0
Wayland · ·	1,535	*Decrease of 1s.*	11	0	10	0
CARNARVONSHIRE— Carnarvon · · (Gwyrfai Rural Dist.).	1,124†	Labourers without food, advance of 1s.	19	0	20	0
		Labourers with food, advance of 1s.	11	0	12	0

* Agricultural labourers in this district are hired in March and April for a year certain, and the change noted applies to the whole year, and not to the summer only.

† The number of agricultural labourers, etc., is for the Poor Law Union, but the change applies to the Rural District only.

‡ This number is partly estimated.

The adjoining tables give examples of the way in which the changes in agricultural wages were tabulated in the Second and Third Report on Changes in Rates of Wages and Hours of Labour. In the first table space is wasted by devoting separate columns to increases

Agricultural wages : Change in tabulation.

* From the third *Annual Report on Changes of Wages*, pp. 118, 119, 121 (typography adapted).

and decreases, with the intention of making the table distinct; while it is not clear whether " Winter 1894 " means the winter beginning in or that ending in that year.

In the second table, which refers to summer wages only, the columns are rearranged; and increases and decreases printed in the same column, the latter in italics. In the Fifth Report all the information is printed in a clearer way, thus :—

WINTER WAGES.*

District.	Number.	Weekly Rates.		Increase or Decrease per Week in 1897.	
		Jan. '96.	Jan. '97.	Increase.	Decrease.
		s. *d.*	*s.* *d.*	*s.* *d.*	
Tendring	3,113	10 0	11 0	1 0	...

The tabulation is repeated for the summer.

The weakness in these agricultural returns is in the numbers column. In the returns from other industries the numbers given are those actually affected, but in this case it is not found possible to obtain this number correctly, and the number entered is that found under " agricultural labourers " in the 1891 census, which includes the various categories as given in the above table. When a change of wages takes place in a rural district, we may perhaps assume that it is likely to be general, though, if it was a reduction, it might not be made by the better employers; and though the change will not take place in the same week throughout the district, there is not likely to be much variation in this respect. The change was generally made at the time that winter wages gave place to summer, or summer to winter; and a slight increase or decrease may take place by making the winter reduction or the summer advance later than usual. On the whole, little error will be introduced by assuming that the change stated affects all the adult agricultural labourers in the district, and it is quite probable that a proportional change †
will take place in the wages of horsekeepers, shepherds, and others, though it may not in the case of boys, or old men who are earning less than the district rate. The question,

The number affected.

---

* From the fifth *Annual Report on Changes of Wages*, p. 145.
† On these points see Mr. Wilson Fox's Report on *Wages and Earnings of Agricultural Labourers*, 1900, p. 50, and pp. 111-157.

" Approximate number of able-bodied labourers in parish ? "
is asked on the inquiry form, but as the answers are not used,
it may be assumed that they are generally not given with
sufficient exactness.

The object of the whole tabulation is to show the change in
the national weekly wages bill, but many details are lacking for
the complete calculation.   In the case of agricultural labourers,
we need, in addition to these data, accurate statements of the
Lack of data. change of additional earnings, special payments,
and payment in kinds.   In all cases we need a
more complete account of the whole wage-bill as well as the
change.   For agricultural labourers the material has been
published by the Labour Department ; * every year it received
returns from most of the 600 unions as to wages at all seasons,
whether there has been a change or not.

The looseness in the returns as to numbers does not prevent
our calculating the change in the county or country rates, for
Changes in county rates. the numbers in each district affected by the
change may be expected to bear the same pro-
portion to the numbers given in the census returns, as the
number of agricultural labourers of the same class in the
whole county or country does to the census number, and we
are helped by the principles of weighted averages discussed
in the next chapter.

The calculation for Durham in the above table for the
changes in summer wages 1894–95 may be performed as
follows :—

	Average before change.	Change.	Proportional number affected.	Amount of change on wage-bill.
	s. d.			*s. d.*
Stockton · ·	17 6	– 6d.	4	– 2 0
Teesdale · ·	17 6	+ 6d.	7	+ 3 6

Total change in county, + 1s. 6d.
Proportional number in county, 73.

Effect on county average, $\dfrac{1/6}{73} = \frac{1}{4}$d.

Here, for simplicity of calculation, the numbers affected are
taken to the nearest 100, a process which is not likely to affect

---

* On these points see Mr. Wilson Fox's Report on *Wages and Earnings
of Agricultural Labourers*, 1900, p. 50, and pp. 111–157.

the average perceptibly.* This rough method is likely to give the result as accurately as the original data make possible. A similar process with suitable modifications can be applied to the changes tabulated for other industries. The summary of such returns for agriculture for all counties is as follows :—

COMPARISON OF THE NET EFFECT OF THE CHANGES OF CASH WAGES
  • per Week paid in the Years 1896 and 1895 in certain Districts in England and Wales.†

DISTRICT.	WAGES IN 1896 AS COMPARED WITH 1895.			WAGES IN 1895 AS COMPARED WITH 1894.		
	Total ** Number.	Net Effect of Changes on Weekly Wages. Increase (+) and Decrease (−).		Total ** Number.	Net Effect of Changes on Weekly Wages. Increase (+) and Decrease (−).	
		Total.	Per Head.		Total.	Per Head.
ENGLAND—		£	s. d.		£	d.
Northern Counties -	5,662	− 43	− 0 1¾	3,766	+ 44	+ 2¾
Yorkshire, Lanca-shire, and Cheshire	2,897	+ 100	+ 0 8¼	3,942	− 126	− 7¾
Eastern and Midland Counties - -	69,869	+ 666	+ 0 2¼	89,576	− 2,045	− 5½
Southern and Wes-tern Counties -	20,901	− 340	− 0 4	20,441	− 575	− 6¾
WALES - - -	...	...	...	2,165	+ 73	+ 8¼
Total - -	99,329	+ 383	+ 0 1	119,890	− 2,629	− 5¼

** The number given is the total of male agricultural labourers, farm servants, shepherds, horse-keepers, in 1891, in the Poor Law Unions in which the changes took place.

* The corresponding calculations for Oxfordshire are :—

12/	−1/	11	−11/
13/	−1/	16	−16/
			−27/

Effect on county average, $\dfrac{-27/}{161} = -2d.$

For Norfolk :—

12/	−1/	134	−134/

Effect on county average, $\dfrac{-134/}{425} = -4d.$

† From the fourth *Annual Report on Changes of Wages*, p. xliv.

The value of this table is not obvious. It seems of little importance to know how many persons were affected alto-
Criticism of summary table. gether; though it is of some value to learn from a previous table that 58,578 persons received increases, and 40,751 decreases in 1896. This total of persons affected is constantly given in these tables; if a person receives an increase of 1s. one month, and loses it the next, he is counted as 2, and his contribution to the next column (net effect, of change) is zero. This — £43 may mean that 2000 persons received a decrease of 1s. each, and the remaining 3662 (same or different persons) an increase of $3\frac{3}{4}d$. each, or any other figures which would give the same total. The change per head in the next column is unimportant; it only shows an arithmetical quotient with no concrete meaning that can be expressed in words. If it was replaced by another quotient, viz., $\frac{£43}{n}$, where n is the number of agricultural labourers in the Northern Counties, we should know the effect on average wages. In fact, the table would be more useful thus :—

APPROXIMATE EFFECT OF CHANGES ON NATIONAL WEEKLY
WAGE BILL.

DISTRICT.	INCREASES.		DECREASES.		Net Change.	Total No. Employed.	Average Change.
	No. affected.	Total.	No. affected.	Total.			

The figures given supply an example of the common practice of carrying out into detail a calculation which depends originally on incorrect numbers, in this case the number employed, and is therefore misleading throughout. Till the average (useless here in any case) is taken, the error in this quantity has no injurious effect. As shown above, the average here given could be replaced by another which would be of use, and which would be correct within limits that could be defined, and would be narrow enough for most purposes.

Further, since the column of numbers affected is admittedly wrong, the figures should be given to the nearest 1000 rather

than to units, even if no attempt was made to estimate the new figure; " between 5000 and 6000 are affected." is a more useful and correct statement than " 5662 persons belonged in 1891 to a class in some undefined way connected with that in question in 1896."

Since the introduction of the minimum wage in agriculture the whole problem has been modified and simplified. The foregoing analysis, however, still illustrates the adaptation of tabular methods to difficult and imperfect data, and shows how records of wage-changes in general were handled officially for at any rate twenty years.

The discussion of Group C, the tabulation of non-numerical answers, must be postponed till we have analysed the nature and use of averages.

CHAPTER V.

AVERAGES.

IT is natural, in a book with the present title, to allot a considerable space to averages. By the use of averages complex groups and large numbers are presented in a few significant words or figures; and thus the two definitions of statistics, the *Science of Averages* and the *Science of Large Numbers*, are reconciled.

Some writers have attempted to draw a distinction between *averages* and *means*, but no general agreement has been reached
Averages and means. as to the exact senses in which the words are to be separately applied.* The best distinction may be made by deciding that an average is a purely arithmetical conception, such as the average length of life in a varied population, which does not correspond to any particular group, but is only a short way of expressing an arithmetical result; while the word " mean " is to be applied to some objective quantity, such as the mean height of Englishmen, about which all height-measurements are grouped in a definite way. If this terminology is adopted, most of the discussion under A, B, C in the sequel applies to " averages " and under D, E and F to " means."

A. ARITHMETIC AVERAGES.—We may rapidly pass by some of the common uses of the word " average," and pick out those which will prove of use in statistics. An average is sometimes used merely to avoid big numbers. The average weight of the University crew is given, only because it is more usual to speak of a man's weight being $12\frac{1}{2}$ stone than of eight men's weight being $12\frac{1}{2}$ cwt., and it is easier to connect the former with men's weight in general. Similarly, if we are comparing the value of the exportations of some commodity

---

* Compare the article "Moyenne," by, Dr. Bertillon, in *Dictionnaire encyclopédique des Sciences Médicales,* with this chapter. See also the paper by Dr. Venn in the *Statistical Journal,* 1891, and chap. xviii. in his *Logic of Chance.*

in two periods of ten years each, we should say that the yearly average in the period 1870–79 was £10,000,000, and 1880–89 was £11,000,000, rather than that the totals were £100,000,000 and £110,000,000. This leads to the second ordinary use of the word. If we were comparing the ten years　The common 1870–79 with the eleven years 1880–90, and the　denominator. totals in the periods were £100,000,000 and £132,000,000 respectively, we should obtain no grasp of the difference till we had reduced them to a common denominator by dividing by the number of years, and found that the averages in the two periods were £10,000,000 and £12,000,000. This class of averages is well known in cricket; sometimes the total number of runs made or wickets taken by each cricketer are stated also, but these are rather as so-called statistical curiosities than as having much bearing on the skill or luck of the players. The numbers by which the seasons' performances are judged are the quotients of the number of runs by the number of innings, of the number of wickets by the number of runs, and so on, all quantities being reduced to a common denominator. The average in this sense is very common in mechanics. The average pressure per square inch, the average work done by an engine per minute, the average speed of a train, are quantities which it is frequently necessary to use. Such an expression as the average rate of interest is precisely similar.

It will be clear that percentage is a special case of this use of average. It is useless when comparing the growths of population or of trade to give only the whole　Averages as numbers. An increase of 50,000 in the population　rates. of London is not so significant as one of 10,000 in that of Harrow; they must be expressed as increases of 1 per cent. and 60 per cent., say, before their meaning can be appreciated, and this is the same thing as giving the average increase to 100 inhabitants. For this reason the records of births, deaths, and marriages are always given as rates—so many per 1000 inhabitants; and in these cases a double average is given, for the rates signify so many per 1000 inhabitants per annum.

Another extension of the same use is found when quantities are reduced to rates " per head " of the population. This use is solely for comparison, and the principle employed is that of the common denominator. It would be futile to state that the amount spent on drink was, say, £100,000,000 in

1860 and £110,000,000 in 1890; but the corresponding state-ments that the amounts were £3 10s. per head in 1860 and £2 15s. per head in 1890 would make a comparison possible. In preparing any comparative summary of figures, it is always necessary to consider whether such an average should be taken.

Preliminary definition. So far, the averages considered are simply arithmetical, and satisfy the following definition :—

Average × number to which it applies = total quantity dealt with.'

e.g. Average weight × number of crew = total weight of crew.

The following question, however, will lead us further. The average weekly agricultural wages in 1892 in Wilts, Dorset, Devon, Cornwall, and Somerset were 10s., 10s., 13s. 6d., 14s., 11s. respectively. What was the average in the south-west of England?

Its inapplica-bility.

The simplest method is to say, the average was

$$\frac{10s.+10s.+13s.\ 6d.+14s.+11s.}{5}=\frac{58s.\ 6d.}{5}=11s.\ 8{\cdot}4d.$$

and for many purposes this would be sufficient; but it does not satisfy the above definition. For when we ask the double question " 11s. 8·4d. multiplied by what number equals what total ? ", we can only answer that 11s. 8·4d. multiplied by the number of items equals the sum of items.

We must consider further what we understand by the expressions " average wage in each county," and " average wage in the group of five counties."

It may be supposed that the average wage in Wilts, for instance, was compiled by getting returns from different villages, say 12s., 11s., 9s., 9s. 6d., 10s. 6d., 9s., 9s., adding them and dividing by the number of villages. This of course satisfies our definition no better than the former. What is to be understood by the average in each village ? If our present definition is to be satisfied, it should be the total of the wages paid in the village divided by the number of workers. It is hardly necessary to say that this total is never found in such an investigation, and the average is given from observation · or by guess-work, not by calculation.

If, however, the village average was correct, and we had returns from all the villages in the county, we should find the county average as follows :—

$$\frac{12/\times200+11/\times150+9/\times300+9/6\times150+10/6\times400+9/\times200+9/\times200}{200+150+300+150+400+200+200}=9/11{\cdot}8$$

where the numbers in the denominator are the numbers of labourers in the respective villages. We should then have the same result as if we had had the wages of all the labourers in the county put down on a sheet, added up, and divided by their number, and the average would satisfy the definition.

It is clear that we can simplify this arithmetical work, for if we divide throughout by 50 we get the same result; this is as if we said there were 4, 3, 6 . . . labourers in the villages instead of 200, 150, . . . Thus we get the same result if we take numbers proportional to the total numbers of the labourers instead of the actual numbers. This plan has two advantages : first, that though we do not know the numbers of labourers, we know numbers nearly proportional to them, viz., those included in the census returns under the general headings relating to agriculture; and secondly, we need not choose our numbers with absolute exactness; thus the numbers of labourers above given may be supposed to be round numbers substituted for 213, 145, 320 . . .; and it will presently be seen that such differences hardly affect the average. We idealise the village, and suppose it to contain round numbers; and then for the numerical work take simple numbers proportional to these. This is important as simplifying numerical work.

Averages obtained for the county in this way do not absolutely satisfy our definition, but are very nearly equal to those that do. We can then proceed to take the average for the south-west of England on the same principles.

A common case is when the data are given as so many instances in successive grades, as in columns 1 and 3 in the following table. To obtain the arithmetic average it is necessary to make some assumption as to the distribution of the instances within the grade. It can be shown that in ordinary cases, especially where the Graded data. numbers tail off rapidly at both extremities, a high degree of accuracy is obtained by setting out the work as if the numbers ·in each grade were concentrated at the middle point of that grade; in fact the average in each grade is generally nearer the centre of the group than is the middle point of that grade, but the resulting errors on either side of the centre tend to neutralise each other. The work is generally simplified by taking the breadth of the grade (five years in the table) as

unit, and measuring from an origin selected at the middle point of a grade in which the entries are numerous (grade 40–45 years, middle point 42½ in the table). The average distance from this origin (obtained by dividing the total of column 4 by the number of cases) shows the distance of the average from the origin in terms of the unit, whence the average is readily calculated on the original scale.

AGES OF MARRIED MEN IN ENGLAND AND WALES, 1911.

Grade Years.	Middle Points of Grade measured from Origin, 42½ Years; Unit, 5 Years.	Number of Married Men per 1000.	Product of Numbers in Cols. 2 and 3.	Cumulation.	
				Limiting Age.	Number above Age in Col. 5.
Col. 1.	Col. 2.	Col. 3.	Col. 4.	Col. 5.	Col. 6.
15–20	−5	0	0	15	1,000
20–25	−4	33	−132	20	1,000
25–30	−3	112	−336	25	967
30–35	−2	152	−304	30	855
35–40	−1	154	−154	35	703
40–45	0	136	0	40	549
45–50	1	118	+118	45	413
50–55	2	96	192	50	295
55–60	3	74	222	55	199
60–65	4	54	216	60	125●
65–70	5	37	185	65	71
70–75	6	21	126	70	34
75–80	7	9	63	75	13
80–85	8	3	24	80	4
85–90	9	1	9	85	1
90–95	10	0	0	90	0
		1,000	+1,155 − 926		
			229		

Average : $42\frac{1}{2} + \frac{229}{1000}$ of $5 = 43 \cdot 645$ years.

Diagrams illustrating this table are given facing p. 130.

B. WEIGHTED AVERAGES.—This discussion introduces and gives an example of the very important statistical method known as "weighting the average." We may illustrate it further from the same figures by considering what weights to apply to get this average for South-West England. We may find the number of agricultural labourers in the counties and work out the average thus : $\dfrac{10s. \times 20,000 + 10s. \times 30,000 +}{20,000 + 30,000 +}$; or we may argue that since we have no means of knowing the

exact numbers of labourers we may as well arrange the weights, according to the importance of the counties, say 20,000, 30,000, etc., from some other point of view, and take numbers representing such quantities as the amounts of wheat produced, the area, or the rate of increase of population. In this particular case these methods would be absurd, but in other problems the weights are not so obvious. Suppose, for example, that we are considering the attraction of London on the inhabitants of various counties; that we are told that so many immigrants arrive from Essex, Norfolk, and Suffolk, and so many from Stafford and Worcester, and we are asked to compare the attractive power on the agricultural and manufacturing counties. Should we weight the numbers given by the total numbers of inhabitants of the contributing counties, or by their distance from London, or by some quantity derived from these?

The idea is made clearer by the mechanical analogy in which the word *weight* originated. Suppose a uniform weightless rigid rod graduated in 100 equal divisions, Mechanical and equal weights hung at the 40th, 50th, 60th, illustration. 70th, and 80th divisions from one end; the rod will then balance at a point corresponding to the unweighted average, 60 intervals from the same end. Now, suppose the equal weights replaced by weights of 7, 1, 3, 2, 4 lbs. respectively, and the rod will balance at a point corresponding to the weighted average, 57·1 intervals from the same end. The further any particular mass is moved, or the heavier it is, the more the centre of gravity will be shifted; and this clearly corresponds to the influence we should wish the various wages to have in the statistical problem. The formula in use in Statics $\bar{x} = \frac{Smx}{Sm}$, which corresponds to the arithmetic on the previous page, can also be used in Statistics.

The discussion of the proper weights to be used in this and other averages has occupied a space in statistical literature out of all proportion to its significance, for it may be said at once that no great importance need be attached to the special choice of weights; one of the most convenient facts of statistical theory is that, given certain conditions, the The small effect same result is obtained with sufficient closeness of weights. whatever logical system of weights is applied. We must

postpone the complete mathematical analysis of this proposition, but may offer immediately some algebraic formulæ and arithmetical illustrations.

Write W_1, W_2 . . . W_n for the weights applied to n quantities M_1, M_2 . . . M_n.

Then the weighted average, M_w, $= \dfrac{W_1 M_1 + W_2 M_2 + \ . \ . \ .}{W_1 + W_2 + \ . \ . \ .}$.

Let $\overline{m}$ be the average of the M's, and let $M_1 = \overline{m} + m_1$, $M_2 = \overline{m} + m_2$ . . . .
Then $n\overline{m} = M_1 + M_2 + \ . \ . \ . = n\overline{m} + m_1 + m_2 + \ . \ . \ .$, so that $m_1 + m_2 + \ . \ . \ . = 0$.
Similarly if $\overline{w}$ is the average of the w's, and $W_1 = \overline{w} + w_1$, etc., $n\overline{w} = W_1 + W_2 + \ . \ . \ .$, and $w_1 + w_2 + \ . \ . \ . = 0$.
Then $(W_1 + W_2 + \ . \ . \ .) M_w = (\overline{w} + w_1)(\overline{m} + m_1) + (\overline{w} + w_2)(\overline{m} + m_2) + \ . \ . \ . ;$
$\therefore n\overline{w} . M_w = n\overline{w}\,\overline{m} + \overline{m}(w_1 + w_2 + \ . \ . \ .) + \overline{w}(m_1 + m_2 + \ . \ . \ .)$
$\qquad\qquad\qquad\qquad\qquad\qquad + w_1 m_1 + w_2 m_2 + \ . \ . \ . ;$
$\therefore M_w - \overline{m} = \dfrac{1}{n}\left\{ \dfrac{w_1}{\overline{w}} m_1 + \dfrac{w_2}{\overline{w}} m_2 + \ . \ . \ . \right\}.$

The difference between the weighted average (M_w) and the unweighted average $\overline{m}$ depends therefore on the average of terms such as $\dfrac{w_1}{\overline{w}} . m_1$. The sum of the w's and the sum of the m's is zero, and of the w's and of the m's many are negative and many positive. It is only when like signs are more commonly found in a pair of m and w than are unlike signs that the whole expression for the difference between M_w and $\overline{m}$ becomes at all important.

In the following table from the Wage Census (see page facing), $\overline{m}$ is 24s. 2d., $M_w = 24s$. 7d., $n = 38$, $\overline{w} = 94^{00}$, and writing the weights to the nearest 100 and the wages in pence we have the following values, the trades being taken in the order of the table :—

w.	m.	wm.	w.	m.	wm.
+228	+13	+2,964	+433	+ 41	+17,753
+ 28	−12	− 336	+149	− 41	− 6,109
− 24	−10	+ 240	+186	+ 36	+ 6,696
− 26	−53	+1,378	− 42	+ 7	− 294
− 66	−58	+3,828	− 32	+ 4	− 128
− 82	− 8	+ 656	+323	+ 19	+ 6,137
− 72	−23	+1,656	+ 13	+ 61	+ 793
− 81	+29	−2,349	− 79	+111	− 8,769
− 83	+ 3	− 249	− 73	+ 1	− 73
− 88	+37	−3,256	− 76	+ 65	− 4,940
− 67	−48	+3,216	− 89	+ 50	− 4,450
− 91	−36	+3,276	− 91	+ 75	− 6,825
+580	−15	−8,709	− 77	+ 28	− 2,156
− 44	−92	+4,048	− 65	+ 1	− 65
− 64	+10	− 640	− 10	+ 1	− 10
− 25	−25	+ 625	− 76	− 46	+ 3,496
− 71	−27	+1,917	− 62	− 16	+ 992
− 54	− 4	+ 216	− 83	− 14	+ 1,162
− 89	−66	+5,874	− 72	+ 12	− 864

Sum of 21 positive products +66,923.
Sum of 17 negative products −50,213.
Sum of the 38 products = 16,710 = $w_1 m_1 + \ . \ . \ .$
$M_w = \overline{m} + \dfrac{1}{n\overline{w}}$ of 16,710 = 24s. 2d. $+ \dfrac{16,710}{38 \times 94} d. = 24s$. 6·68d. = 24s. 7d. to nearest penny, as in the table.

The table on the next page affords an example of this

EXAMPLES OF THE SMALLNESS OF THE CHANGE INTRODUCED BY DIFFERENCE IN SYSTEMS OF WEIGHTING.

Trade.	Average Wages (Men).		Number Included in Returns	Numbers Employed in Trade when known. Unit 1,000	Arbitrary System of Weights.	Equal Weights.
	s.	*d.*				
Cotton Manufacture - - -	25	3	32,189	142	144	1
Woollen „ - - -	23	2	12,248	54	172	1
Worsted and Stuff Manufacture -	23	4	7,005	38	219	1
Linen Manufacture - - -	19	9	6,807	22	96	1
Jute „ - - -	19	4	2,799	9	23	1
Hemp, &c., „ - - -	23	6	1,232	3	78	1
Silk „ - - -	22	3	2,248	10	189	1
Carpet „ - - -	26	7	1,292	0	213	1
Hosiery „ - - -	24	5	1,070	8	287	1
Lace „ - - -	27	3	593	8	51	1
Smallwares „ - - -	20	2	2,734	0	225	1
Flock and Shoddy Manufacture -	21	2	330	2	200	1
Coal, Iron Ore, and Ironstone Mines - - - -	22	11	67,429	57	142	1
Metalliferous Mines - - -	16	6	5,046	0	190	1
Shale Mines and Paraffin Oil Works	25	0	3,021	0	207	1
Slate Mines and Quarries - -	22	1	6,933		232	1
Granite Quarries and Works -	21	11	2,315	12	206	1
Stone Quarries - - - -	23	10	3,956		34	1
China, Clay, &c., Works - -	18	8	499	0	39	1
Police - - - - - -	27	7	52,682	58	224	1
Roads, Pavements, and Sewers -	20	9	24,276	0	29	1
Gasworks - - - - -	27	2	27,965	0	40	1
Waterworks - - - -	24	9	5,187	0	151	1
Pig Iron (Blast Furnaces) - -	24	6	6,234	0	128	1
General Engineering Iron and Brass Foundries and Machinery Trades - - - - -	25	9	41,658	200	173	1
Shipbuilding, Iron and Steel -	29	3	10,661	80	228	1
Tinplate Works - - - -	33	5	1,514	0	178	1
Saw Mills - - - - -	24	3	2,088	0	174	1
Brass Works and Metal Wares -	29	7	1,838	0	222	1
Shipbuilding, Wood - - -	28	4	454	0	79	1
Cooperage Works - - -	30	5	327	0	165	1
Coach and Carriage Building -	26	6	1,664	0	28	1
Boot and Shoe Making - -	24	3	2,902	0	142	1
Breweries - - - - -	24	3	8,366	0	46	1
Distilleries - - - - -	20	4	1,795	0	129	1
Brick and Tile, &c., Making -	22	10	3,188	0	55	1
Chemical Manure Works - -	23	0	1,054	0	210	1
Railway Carriage and Wagon Building - - - -	25	2	2,239	0	233	1
			s. d.	*s. d.*	*s. d.*	*s. d.*
Averages - -	...		24 7	25 3	24 5½	24 2

principle,* and is worth careful study.   At the commencement
Example from of the Wage Census, circulars were sent to all the
the Wage Census. principal firms in all well-located trades, asking
for details as to wages.   Of these some were not returned,
and the numbers allotted in the Final Report to each trade
are not the numbers which actually belong to the trade in the
whole country, but the numbers of those in the firms which
made returns.   The average wage given is not therefore the
arithmetic average for these trades for the whole country
corresponding to the definition given above for average, but
the average of the average wages as returned in each trade
weighted by the numbers for whom returns were made; so
that the average wage given for the whole group of trades
might have proved to be different, if with the same average in
each trade the returns had been complete.   It is very unlikely,
however, that there would have been any great difference.
In the table several systems of weighting are used; the first
are the numbers in these returns, giving an average, 24s. 7d.;
the second are the numbers belonging to each trade according
to the census when they are above a certain minimum, giving
an average 25s. 3d.; the third is a purely arbitrary list of
figures taken from a source which has no connection with
wages, and the average is 24s. 5½d.; the last is the unweighted
average, that is, all the weights are equal, and the average is
now 24s. 2d.   These averages are close together, while the
original items vary from 16s. 6d. to 30s. 5d.   It is to be noticed
that the true weights are not known in this case, but that
owing to this principle we are able to dispense with them
entirely.

The problem dealt with in the next table is to find the
average weekly agricultural wage in England and Wales from
Uniformity of the returns for Michaelmas 1869 and Lady Day
average under 1870, given in columns 1 and 2.   There are very
many systems many different ways of taking this average, some
of weights. of which are as follows :—Take the average of
summer and autumn for each county, as in column 3, and then
the unweighted average of these 45 numbers; this is 12s. 7d.
Suppose the summer wage to be paid twice as long as the
autumn wage, as in column 4, and proceed as before; the

---

average' is 12s. 5½d., the slight difference being due to the inclusion of harvest payments in the Michaelmas wage, which makes them higher on the whole than the summer wages. Again, divide the counties into geographical groups, take the simple average for each group (the figures marked *a* in column 3 and *b* in column 4) and weight these by the figures marked *c* in column 5, the numbers of agricultural labourers in each group; the average of the *a* figures with the *c* weights is 12s. 5d., of the *b* figures with the *c* weights is 12s. 4d. Again, weight the figures for each county in column 4 with the numbers in column 5, the most obvious method of all; the average is then 12s. 4d. ' Again, take the simple average of the district averages *a* and *b*, that is, give each of the eight districts equal weights; the averages are 12s. 4¾d. and 12s. 3¼d. Or take the simple average of column 3, counting Yorkshire and Wales each as one county; it is 12s. 8d.

To obtain new groups, take as weights not the number of agricultural labourers, but the total population of the districts, the numbers marked *d*. Exclude the population of London as exerting a preponderating influence unconnected with agriculture. A new factor is now introduced, for population is greatest in the manufacturing districts, where agricultural labour is of comparatively little importance, but receives high wages; these high wages have undue weight, and the average of the figures *b* with weights *d* is brought up to 13s. 1¾d. If column 4 is rewritten correct only to the nearest 1s., and column 5 to the nearest 10,000, the weighted average is 12s. 5d. If column 3 is weighted with random numbers quite unconnected with the problem, viz., the successive digits in the third decimal places of the logarithms of the numbers 2 to 46, the average is 12s. 10¾d. The reader may try any other system of logical or absurd weights, and he will find that unless there is some bias in the selection of weights, or great preponderance is given to a few counties, that the average will be little affected.

Since the true system of weights which would reduce the general average to our definition must be allied to some of those here adopted, and can hardly show greater divergence from 12s. 4d. than these do, we may feel confident that the true average is within, say, 3d. of this figure. The original items varied from 8s. 6d. to 19s.; the averages, even those based on the most extravagant methods, are contained by the

AGRICULTURAL WAGES IN 1869-70. To Illustrate Various Methods of Weighting, and their Results.

	1. Michaelmas 1869.	2. Lady Day 1870.	3. Average of Cols. 1 and 2.	4. Average of Col. 2×2 and Col. 1.	5. No. of Agricultural Labourers in Counties. Unit 1,000.	6. Whole Population in Groups of Counties. Unit 100,000.
	s. d.	s. d.	s. d.	s. d.		
Sussex	12 3	12 6	12 1½	12 1	34	...
Surrey	14 0	13 6	13 9	13 3	16	...
Kent	14 6	14 6	14 3	14 2	44	...
Hants	11 0	10 6	10 9	10 8	32	...
Berks	12 0	12 0	12 0	10 8	22	...
Average	...	...	a 12	b 12 3	c 148	d 22
Herts	14 6	11 10	13 2¾	12 9	20	...
Northants	12 6	11 6	12 0	11 10	23	...
Hunts	16 0	11 0	13 6	12 8	9	...
Bedford	13 0	11 0	12 6	12 4	17	...
Camb.	11 0	12 6	11 6	11 8	24	...
Average	...	...	a 12	b 12 3	c 93	d 14
Essex	12 6	11 0	11 9	11 6	45	...
Suffolk	10 6	11 6	10 9	10 10	41	...
Norfolk	11 6	11 0	11 6	11 6	44	...
Average	...	...	a 11	b 11 3	c 130	d 12
Wilts	11 0	10 3	10 7½	10 6	26	...
Dorset	9 6	10 3	9 10¾	10 0	17	...
Devon	10 0	10 3	10 1½	10 2	34	...
Cornwall	11 0	11 6	11 0	11 8	17	...
Som. &c.	11 0	10 6	10 9	10 8	31	...
Average	...	...	a 10	b 10 6	c 125	d 19

	1. Michaelmas 1869.	2. Lady Day 1870.	3. Average of Cols. 1 and 2.	4. Average of Col. 2×2 and Col. 1.	5. No. of Agricultural Labourers in Counties. Unit 1,000.	6. Whole Population in Groups of Counties. Unit 100,000.
	s. d.	s. d.	s. d.	s. d.		
Stafford	13 0	13 0	13 3	13 7	19	...
Gloucester	11 0	10 0	10 6	11 1	22	...
Hereford	10 3	10 0	10 3	11 1	12	...
Salop	11 6	11 0	11 3	11 4	21	...
Worcester	13 6	11 0	13 3	11 6	15	...
Warwick	13 6	12 0	12 9	12 6	20	...
Average	...	...	a 11 9	b 11 7	c 109	d 27
Leicester	14 0	13 0	13 6	13 4	15	...
Rutland	12 6	12 6	13 6	12 8	3	...
Lincoln	14 6	13 0	13 9	13 8	49	...
Notts	13 6	13 0	13 9	13 2	16	...
Derby	13 6	14 0	13 9	13 10	8	...
Average	...	...	a 13 3½	b 13 3½	c 91	d 14
Cheshire	13 6	13 0	13 6	13 3	18	...
Lancs.	15 0	15 0	15 0	15 6	30	...
Yorks, W.	19 4	15 3	17 1½	16 9½	30	...
Yorks, N.	16 6	13 6	16 5	14 2	16	...
Durham	16 6	16 6	18 0	16 6	8	...
Northumberland	19 6	16 6	16 0	17 6	1	...
Cumberland	15 3	15 0	18 0	15 0	10	...
Westmoreland	16 3	15 6	15 10½	15 9	3	...
Average	...	...	a 15 9	b 15	c 127	d 72
Monmouth	12 6	13 6	13 1½	13 4	6	...
Glamorgan	14 6	14 6	14 11	14 6	5	...
Caermarthen	12 4	14 6	14 11	11 11	4	9
Pembroke	11 0	10 0	10 6	10 4	4	...
Cardigan	9 0	8 0	8 0	8 8	5	...
Brecknock	12 0	12 0	12 0	12 0	4	...
Radnor	10 0	12 0	10 0	10 0	2	...
Carnarvon	12 0	12 0	12 0	12 0	5	...
Average	...	...	a 11 7	b 11	c 35	d 14

limits 12s. and 13s. 1⅜d. Without some such argument as this we should have no clue to the magnitude of the error introduced by erroneous weights. It is never safe, however, to assume that weights can be neglected, and an unweighted average used, without first examining the group in question, trying various systems, and seeing that the resulting average is stable. This will only be the case if there is no connection between the size of the quantities and the true magnitude of the *Weights cannot in general be neglected, but only the errors in their estimation.* weights. Thus if we are dealing with wages in towns, and are calculating the average for all towns taken together, we shall obtain too small a result if we ignore weights and count all towns as equal, for the higher wages are paid in the larger towns. Thus, as on pp. 118–9 below, the average of the recognised wages of 117 branches of the Amalgamated Society of Engineers was 32s. 4d. in 1891 if we count all the branches as equal; but was up to 33s. 4d. if we weight the wage at each of the branches with the number of members belonging to it. But, though we cannot neglect weights entirely in such cases, we need to make only a very rough estimation for them if there is no preponderating influence exerted by a small minority of places. In this case London, with a wage higher than any other district, except Dartford and Enfield Lock, and with nearly one-sixth of the total number of members dealt with, exerts such an influence. If, giving London its due importance, we take as weights the numbers belonging to the branches to the nearest hundred, we obtain the average 33s. 6d., practically the same as before. Each group for which an average is to be calculated must be treated on its merits; in many cases the weights may be neglected entirely; in nearly all cases, where the group consists of many items, even moderately large errors in computing weights may be neglected. Examination of the data will generally determine the importance of such errors.

This principle is of great importance. In many cases the true weights are incalculable or even undefinable; but now it is seen that, given certain conditions, there is no need to calculate or define the weights; in many other cases the weights cannot be known exactly, but exactness is not necessary. No system of weights, however, can remove an original bias common to all the items. If, for example, wages throughout

were 1s. less than here reckoned, the calculated average would be 1s. too high. So we arrive at a very important precept : *in calculating averages give all care to making the items free from bias, and do not strain after exactness in weighting.*

C. STATISTICAL COEFFICIENTS.—A statistical coefficient is a number, whole or fractional, by which a total (*e.g.*, population) must be multiplied to give an allied number (*e.g.*, number of births). Thus if the birth-rate is 28 per 1000, the coefficient is ·028. These coefficients play an important part in ordinary statistics and a very interesting *rôle* in the application of the law of error to demography. The population may increase or diminish, but the coefficients relating to certain numbers fluctuate within narrow limits and only after a considerable period show any significant change in normal times ; and by their use the statistics of different countries can be compared, and numbers for future years can be forecasted in some cases with marvellous accuracy, subject only to the chance of some great catastrophe. Coefficients can be formed for births (in various districts), for deaths (according to age, profession, or disease), for marriages (at various ages), for suicides, crimes, accidents, consumption of various commodities ; if the preliminary data could be obtained, for the number of persons crossing Westminster Bridge in the year, the number of visitors to the Monument, the number of umbrellas left in the train, and so on ; the list could be prolonged indefinitely. The more important coefficients are calculated for most civilised countries and published in statistical reports. A knowledge of them is necessary for statistical investigations.

It is clear that such coefficients are essentially only a special way of writing a certain class of arithmetic averages, and with reference to them we may discuss more generally the relationship between the terms used on p. 84.

Average (A) × number to which it applies (N) = total quantity (Q) dealt with,

or $$A = \frac{Q}{N}, Q = N \times A.$$

Thus in the case of births A is the coefficient, N the population, Q the number of births.

So far as is practicable, a movement of Q should reflect change in only one factor. If N is the whole population, Q

will be affected by changes in the sex and age distribution of the population, and by the number of marriages and age at marriage, as well as by fecundity. Methods of securing strict comparability between the denominators in the cases of birth, marriage and death rates (by means of correcting factors *) are in common use. When these methods are not applicable we may fall back on the rule given by Bertillon (*Cours élémentaire*, pp. 94 *seq.*), effects (Q) should be compared with their immediately productive causes (N); thus in the case of marriages, the question should be put " what persons are capable of marrying ? " and the answer is adult bachelors or spinsters or widowers or widows, and the total of these groups gives N. The rule may be extended to include persons or things indirectly concerned or affected; thus the output of coal may be considered in relation to coal-hewers (the immediate producers) or to all employed at coal-mines, and the output of domestic coal in relation to the number of private consumers.† To eliminate all factors but one, the entries in the numerator should be homogeneous, the entries in the denominator should be homogeneous, and the potential relation of a person or thing included in the denominator to one in the numerator should be uniform. For example, the average value of exports per head of the population satisfies none of these conditions; exports make a heterogeneous mass, the population consists of both sexes and all ages, and only part of the productive power of the nation is directed to the foreign market.

The crude coefficients and averages, however, have their use; if they change, some factor or factors have changed, and if it is known that all but one are nearly constant, the coefficients move with an identified factor. Thus if N is the population, n the number of marriageable persons, and M the number of marriages, the crude coefficient is given by $C = \dfrac{M}{N} = \dfrac{M}{n} \times \dfrac{n}{N}$; if $\dfrac{n}{N}$ is constant C varies with $\dfrac{M}{n}$, the more logical coefficient.

D. The Mode.—We pass to the consideration of two other means in common use among statisticians but unfortunately

---

* See *Elementary Manual of Statistics* (by the present author), pp. 105-7, and *Statistical Journal*, 1906, pp. 34-147.

† See a discussion on homogeneity, comparability and relativity, *Statistical Journal*, 1908, pp. 463-8.

not yet consciously introduced into common parlance." There are, however, some popular phrases which, if they have any definite meaning, very nearly resemble the averages in ques-

The average man. tion. When we hear of the average clerk, the average working-man, the phrases admit many interpretations. In some way these persons are supposed to be types of their kind. The average clerk may be supposed to mean the one who receives the average income of all clerks, whose expenditure on necessaries and on luxuries is the average of all of his class, who takes the average amount of interest in his work, if of average ability and average age. It will be seen that this clerk is ideal, and not to be found in any random assembly of half-a-dozen; for each of these will have some peculiarity, some quality in which he differs from the average; the average man of the newspapers does not exist in the flesh, but is an imaginary person to whom certain attributes are attached.

Quetelet's average man is familiar; * he is of average height, weight, strength, girth and lung capacity, with eyes of normal

Quetelet's average man. range and medium tint; but he is a more satis- factory model than the newspapers' average, for in regarding him we see the type from which all other men may be supposed to have deviated; the creature that would have been produced if all disturbing causes were removed. That any actual person should answer exactly to all these standards is of course in the highest degree improbable.

Quetelet refers neither to the arithmetic average, nor to the median or the mode (defined in the sequel), but to a mean about which all the similar measurements are grouped in accordance with a definite law, the obedience of anthropo-metrical measurements to which was his chief theme.

The newspaper average, on the other hand, seems to be the mode, the position of the greatest density, which may be

The mode. explained as follows :—Referring back to the table of American wages, p. 69, or to the table on next page, it will be noticed that in looking down column 2 we find the numbers increase till we come to 685 (between $1.15 and $1.24), and then after fluctuations diminish. This number, 685, is the greatest which occurs in any 10-cent group.

---

* See Quetelet's *Physique Sociale ;* and Edgeworth in *Statistical Journal,* December 1893.

DETERMINATION OF THE MODE.

Numbers of Wage-Earners from the Senate Report, 1893, *U.S.A*

From $ .25 to		No.	In 20-Cent Groups		In 30-Cent Groups			In 50-Cent Groups	
From $.25 to	.34	1	16		75			317	
" .35 "	.44	15		74		159			
" .45 "	.54	59	144				301		
" .55 "	.64	85		242	355				725
" .65 "	.74	157	270			439			
" .75 "	.84	113		282			483		
" .85 "	.94	169	370		674				
" .95 "	1.04	201		505		1,190			
" 1.05 "	1.14	304	989				1,088	1,472	2,012
" 1.15 "	1.24	685		784	1,242				
" 1.25 "	1.34	99	557			1,023			
" 1.35 "	1.44	458		924			996		
" 1.45 "	1.54	466	538		740				
" 1.55 "	1.64	72		274		603			934
" 1.65 "	1.74	202	531				589	1,297	
" 1.75 "	1.84	329		387	660			970	
" 1.85 "	1.94	58	331			376			
" 1.95 "	2.04	273		418			583		
" 2.05 "	2.14	45	310		343				640
" 2.15 "	2.24	265		298		399			
" 2.25 "	2.34	33	134				330		
" 2.35 "	2.44	101		297	310				
" 2.45 "	2.54	196	209			372		506	
" 2.55 "	2.64	13		176			178		322
" 2.65 "	2.74	163	165		180				
" 2.75 "	2.84	2		17		146		198	
" 2.85 "	2.94	15	144				149		
" 2.95 "	3.04	129		134	181				
" 3.05 "	3.14	5	52			64			285
" 3.15 "	3.24	47		59			59		
" 3.25 "	3.34	12	12		233			254	
" 3.35 "	3.44	0		221		226			
" 3.45 "	3.54	221	226				242		
" 3.55 "	3.64	5		21	32				114
" 3.65 "	3.74	16	27			27			
" 3.75 "	3.84	11		11			93		
" 3.85 "	3.94	0	82		82			96	
" 3.95 "	4.04	82		82		85			
" 4.05 "	4.14	0	3				3		6
" 4.15 "	4.24	3		3	3				
" 4.25 "	4.34	0	0			3		4	
" 4.35 "	4.44	0		3			4		
" 4.45 "	4.54	3	4		4				
" 4.55 "	4.64	1		1		1			9
" 4.65 "	4.74	0	0				0		
" 4.75 "	4.84	0		0	8				
" 4.85 "	4.94	0	8			8		8	
" 4.95 "	5.04	8		8			8		
" 5.05 "	5.14	0	0		1				1
" 5.15 "	5.24	0		1					
" 5.25 "	5.34	1	1					1	

The value of the graded quantity in a statistical group (of wages, heights or some other measurable quantity) at which the numbers registered are most numerous is called the *mode,* or the position of *greatest density,* or the *predominant value.* In the case of a group that is represented by a continuous curve the value is the abscissa of the *maximum ordinate.*

In this column 2 we have, however, 14 maxima in the correct sense of the word, the numbers rise and fall with little

Method of determining the mode. regularity, and there are 14 modes of which that at $1.15–$1.24 is the most pronounced. But if the groups are made wider, and the numbers entered as in column 6 in half-dollar limits, there are only three modes, or if we neglect the small group of 8 at $5.00 only two. The position of the largest group of 1472 is not at once assignable more closely than as between .75 and 1.25.

A further method of approximating to the mode may be illustrated as follows :—When the numbers are tabulated in 10-cent groups, as on p. 97, the mode is quite indeterminate ; in 20-cent groups the successive numbers beginning at .25–.44 are 16, 144, 270, 370, 989, 557, 538, 531, etc., and the number 989 (in the group $1.05–$1.24) is a distinct mode; if we begin the 20-cent groups at .35–.54, the numbers are 74, 242, 282, 505, 784, 924, 274, etc., and 924 (in the group $1.35–$1.54) is a mode; by this double tabulation it is seen that the 20-cent grouping does not decide the mode. In 30-cent groups we have 355, 674, 1242 ($1.15–$1.44), 740, etc., if we begin with $.55–$.84; we have 439, 1190 ($.95–$1.24), 1023, etc., if we begin with $.65–$.94 ; and 483, 1088 ($1.05–$1.34), 996, etc., if we begin with $.75–$1.04 : the mode by each of these groupings lies in a group which contains $1.15 to $1.24, and this smaller group may be assumed to contain the mode, which is thus at or near $1.20. The example here taken is drawn from a group of very irregular figures, which specially illustrate the difficulties. The method just adopted may be summarised thus :— Tabulate the figures again and again in gradually widening groups till regularity is obtained; then examine again the groups which have the selected width and see if the mode is shifted when the lower limit of the grouping is moved; if it is shifted the groups are not wide enough ; if it is not, the mode is in the smallest group common to the larger equal groups

which all contain it. A diagrammatic method is described on p. 138.

Even when our numbers are initially regular, it is seldom easy to determine the mode exactly. The diffi-culty is best seen by an example. Suppose that we have the following returns as to heights of a large number of men :—

67 in.	-	-	455
$67\frac{1}{4}$,,	-	-	475
$67\frac{1}{2}$,,	-	-	490
$67\frac{3}{4}$,,	-	-	500
68 ,,	-	-	485
$68\frac{1}{4}$,,	-	-	467
$68\frac{1}{2}$,,	-	-	445

At first sight the mode appears to be at $67\frac{3}{4}$ in. exactly; but it must be remembered that even in accurate measurements all heights within $\frac{1}{8}$ in. of $67\frac{3}{4}$ in. will be entered as $67\frac{3}{4}$ if the measurements are taken to the nearest quarter inch, or will have been tabulated in this way if the measurements were more accurate. Hence $67\frac{3}{4}$ in. in reality stands for from $67\frac{5}{8}$ to $67\frac{7}{8}$ in. If the 500 heights so entered were distributed uniformly through this interval, the mode might be given with $67\frac{3}{4}$ in. with fair accuracy; but there are signs in the figures that the mode is below this. Suppose that the figures in reality come from the following measurements :—

From $67\frac{1}{4}$ to $67\frac{3}{8}$ in. 238 ⎫ 483 at $67\frac{3}{8}$ in.
,, $67\frac{3}{8}$,, $67\frac{1}{2}$,, 245 ⎭
,, $67\frac{1}{2}$,, $67\frac{5}{8}$,, 245 ⎫ 495 at $67\frac{5}{8}$,,
,, $67\frac{5}{8}$,, $67\frac{3}{4}$,, 250 ⎭
,, $67\frac{3}{4}$,, $67\frac{7}{8}$,, 250 ⎫ 493 at $67\frac{7}{8}$,,
,, $67\frac{7}{8}$,, 68 ,, 243 ⎭
,, 68 ,, $68\frac{1}{8}$,, 242

and that these had been tabulated as in the last column, the mode would appear as $67\frac{5}{8}$ in.; while the same figures tabu-lated as before gave it as $67\frac{3}{4}$ in. The probability of some such shifting is seen from the original grouping, where the number at $67\frac{1}{2}$ in. is greater than that at 68 in. From this discussion we may see that the mode is always a little indefinite, depending on the width of the groups in which the items are tabulated, on the exact position of the limits of the groups. As the

H 2*

items we deal with become more numerous, we shall find regularity when they are tabulated in narrower groups, and the mode can be assigned with greater accuracy.

A mathematical method (p. 228) suggests that the mode of such a group as given by the heights can be determined by dividing the interval containing the mode ($67\frac{5}{8}$ to $67\frac{7}{8}$ in.) in proportion to the differences between the numbers registered in this interval to the numbers in the adjacent interval, viz. : $500\text{–}490 : 500\text{–}485 = 10 : 15$. The mode so computed is at $\left(67\frac{5}{8} + \dfrac{10}{10 + 15} \text{ of } \frac{1}{4}\right)$ in. $= 67\frac{29}{40}$ in. By this method if two intervals contained the same numbers the mode would be placed at the value dividing them, and if the numbers on either side of that containing the greatest number were equal (if the grouping was symmetrical) the mode would be placed at the centre of the middle interval, in both cases as we should have determined *a priori*.

Now is the " average workman " the man who earns $1.73 per diem, the simple average of the whole group on p. 69, or a The "average man making $1.20 the mode? In ordinary speech man."   the latter is meant. The " average clerk " is not the one whose measurable qualities are an arithmetic mean of all similar qualities, but one whose qualities are found in the same degree in the greatest number of his fellows. There are more clerks who read the evening paper than who read Homer, more who go to music-halls than to oratorios, more whose incomes are £100 than £500, more who live four miles from the City than one or twenty. Even with this explanation the average man is not a real creature, for fortunately no individual has no qualities out of the common. The fact that the average is a pure abstraction is of importance directly we apply statistics to actual affairs; these American workpeople cannot be legislated for in the mass as if they all earned $1.20, or as if those who were alike in this did not differ in other respects, even doing very varying quantities of work for this wage. No single measurement expresses completely even the economic ' Importance of condition of a group of workmen, but if we are the mode.   taking a single measurement, that of the " mode " is often the most useful. It is at the mode that we find the greatest number of whose greatest good we may be thinking. Whereas the arithmetic mean and the " median " (defined

below)'may correspond to no reality but be merely numerical conceptions, the mode is precisely that number for which most instances can be found. It shows the commonest result, that most often obtained, and is of very general application. For an intending passenger by train or 'bus, it is more important to know the most ordinary than to know the average number in a compartment. The mode rather than the average in chest measurements is the number most suitable for the ready-made clothier. For providing a post-office or a store, the mode in postal orders or prices of tea needs to be known rather than any other average. Even the favourite coin in a collection may show the spirit of the congregation better than the arithmetic average of their contributions. In these last instances it may be noticed that the mode is quite definite.

A special feature of the mode is that it is entirely uninfluenced by extremes. A cheque for £1000 in a collection disturbs the arithmetic average, but not the mode. The incomes of a small number of millionaires and an army of paupers may have the same arithmetic average as a nation composed entirely of people moderately well off; but the modes will be very different in the two cases. In considering the change year by year in a group of figures, as for instance, the wages of a large group of workmen, we cannot tell, if we take the arithmetic average as our criterion, whether an improvement is due to a levelling up of the badly paid or a rapid increase for those who were already well off, while the mode will show the changing position of the main body. Mr. Booth's *London* is crowded with instances of the use of the mode. Each age diagram shows the mode in ages for an occupation; each wage list that in wages. His whole description of Class E, the typical workmen of modern towns, is based on the same principle. His measurement of social status, based on the number of rooms occupied or servants employed, can be used easily for stating the mode (four rooms to a family and no servant) but not any other average.

Advantages of the mode.

An objection to this average is that there are many groups of figures to which it is not applicable. If we have a very irregular group of numbers with no particular type, such as the populations of towns in England, the mode would be quite indefinite, and would give no information of importance. The use of the mode is to indicate

Shortcomings of the mode.

the type from which other figures may be regarded as diverging
Thus, in these wage figures, the type is about $1.20, and other
examples lie on either side, wages of men who have for some
reason or other more or less than the normal degree of skill
or opportunity. If there is a type, as in Quetelet's instances,
the mode will show it. The mode only tells us one fact,
however, about each type, and it is necessary to supplement
it with other measurements.

E. THE MEDIAN.—When we are dealing with a group of
persons or things, each of which possesses some measurable
attribute, such as height or wage, we can choose certain quanti-
ties which describe the group in brief. Suppose all the items
arranged in a series in ascending order of the magnitude of
this attribute; the magnitude appertaining to the item half-
way up the series is called the *median*.* Thus if in a group
of wage-earners 200 earn less than 20s. 3d., one earns 20s. 3d.,
and 200 more, 20s. 3d. is the median wage. There are as
many items below 20s. 3d. in the supposed series as above it.
The magnitudes one-quarter and three-quarters up the series
are called the *quartiles ;* * those one, two . . . nine-tenths up
are the *deciles ;* those one, two . . . ninety-nine hundredths
up are the *percentiles*. The median is more definite in position
than the *mode*. When we are dealing with exact measure-
ments, if we have an odd number of items it is the middle one,
if an even number, it lies between the two middle items,
which are in general near together, or coincides with them if
they are equal. If the magnitudes are not given exactly,
but as within small limits, we can by the method described on
pp. 106–7 make a good estimate of their actual values. The
median is not affected by exceptional entries at all; the exist-
ence of any number of millionaires has no more effect on the
median income than of an equal number of any other persons
whose incomes are above the median. For many purposes
it is of course necessary to allow these extreme instances more
weight than those which are nearer the average; but the
arithmetic average often gives them undue weight for this
democratic age, since a single millionaire can counterbalance
thousands of ordinary working men. A further advantage is
that it is extremely simple to find, not needing much arith-

* These quantities have already been used in tabulation, p. 70

metical work, for we need not do more than count those well above and well below the average, and look more carefully at those near it.

• There is a yet more important advantage in the use of the median; it can often be found exactly, when our information as to the items in question is neither accurate nor complete. This will be clear from one or two examples. It may be that in a " wage census" *No need for complete information.* 100,000 persons, whose wages were far below the average, do not come into the returns at all, and it is very difficult to estimate their effect on the arithmetic average for want of information as to their earnings; but to find the median exactly, we need only know their number, not their earnings; and if we can only assign a maximum for their number, we still can place the median within narrow limits. The addition of 100,000 men with wages below 15s. to a general summary for the 356,000 men on p. 470 of the *General Report on Wages in 1886* (C.—6889), would still leave the median in the group 20s. to 25s. where it already is; the change would be very marked, however, in the lower deciles and quartiles, and the arithmetic average would be lowered by at least 2s. 1d. The same argument applies to incomes; information is often very deficient, but it is in many cases possible to assert that a number of men, whose exact income is unknown, receive above a certain assigned sum, or even between two assigned limits, which is all we need to know about them to determine the median, if it lies below the lower limit.

Again, in tracing the history of wages throughout the century it is often very difficult to find the correct average, but at the same time it is frequently possible to say that a very large class of men earned below, say, 15s. a week, and another very large class above 30s. whose wages we do not exactly know, and a more definite number between 15s. and 20s., and 25s. and 30s.; and in order to find the median all we need to do is to investigate more exactly the wages between ·20s. and 25s., if that is the grade which contains it; and even if we have not complete information here, we can still say that the median certainly lies between certain narrow limits. There is yet another advantage, perhaps more important, that the median is applicable to quantities which are *Incommensurable quantities.* not capable of measurement at all. This develop-

ment is especially due to Galton.* Suppose it to be required, for example, to find among a large class of boys the average in intelligence. It is clear that it is not easy to find the arithmetic average of a quantity which cannot be properly measured even by the most elaborate system of marks, but on the other hand it would not be at all difficult with a class of, say, twenty boys, to place them in order of intelligence without committing oneself to such a statement as that *A.'s cleverness was 25 per cent. more than B.'s; and the tenth or eleventh boy in this arrangement would show the style of boys in the class, at least as well as any other average. The disadvantage of this method, the reason why it is not universally applicable,

Disadvantages. is that the median of a series of observations may be totally removed from its type, and in fact may not be situated near any of the different objects which are observed. Thus, if we had two large groups of wages of a thousand men between 15s. and 25s., and another thousand between 35s. and 45s., the median would give us any position between 25s. and 35s., where as a matter of fact not a single wage-earner would be found. The median is then chiefly useful when we are dealing with a series of objects of which the main part lie fairly close together; a few extremes do not affect it.†

If m is the median and a the arithmetic average of n quantities $x_1, x_2 \ldots x_n$, and we call $x_1 - s, x_2 - s \ldots$ the *deviations* of the x's from any quantity s, then m is the value of s which makes the sum of the deviations (all taken positively) a minimum, a is the value which makes the sum of the squares a minimum. The first statement becomes obvious from the following analogy: suppose $2n + 1$ places in a straight line are each served by a single wire from a telephone exchange at the nth place from one end; the lengths of the wires correspond to the deviations; now if the exchange is moved to the $\overline{n + 1}^{\text{th}}$ (or central place), $n + 1$ wires are shortened and n wires lengthened each by the same distance, so that the aggregate of wire is diminished; if the number of places is even, the minimum is obtained at any position at or between the nth and $\overline{n + 1}^{\text{th}}$ from either end. For the second, we notice that $\Sigma x = na$, and that $\Sigma(x - s)^2 = \Sigma x^2 - na^2 + n(a - s)^2$, which is a minimum when $s = a$.

The following table shows the description of 76 items by the help of the various averages now described :—

MEASUREMENTS OF BOYS OF AGES 13 TO 15 YEARS.

No.	Age.	Height.	Weight.	No.	Age.	Height.	Weight.	Tabulation of Weights.
	yrs. mth.	ft. in.	st. lb.		yrs. mth.	ft. in.	st. lb.	
1	14.1	4.11½	6.0¾	39	14.7	4.11½	6.3½	Arithmetic average, 6 st. 1½ lbs.
2	14.9	4.10	5.7	40	13.1	4.11½	5.7	
3	14.7	5.5¼	7.5	41	14.3	4.11	6.4¼	
4	13.11	5.0	6.3¾	42	13.3	4.4½	4.11½	The same, when
5	14.11	5.3¾	8.0½	43	14.3	5.3	6.7½	weights are en-
6	14.7	4.10	5.0	44	13.6	5.1½	6.13¼	tered only to
7	14.3	4.10	6.7	45	14.2	4.8¾	6.0¾	nearest stone,
8	1_.9	5.5	8.5¼	46	13.5	5.2	7.4	6 st. 1¼ lbs.
9	14.11	4.9½	5.12¼	47	13.8	5.2½	6.11	
10	14.13	4.11¾	6.11¾	48	14.6	5.4	7.4½	Median, 6 stones
11	13.4	4.7	5.1½	49	14.8	5.1½	6.10	1⅛ lbs.
12	14.7	5.3¾	7.8¾	50	13.3	4.8¼	5.0	
13	13.8	4.7½	5.3	51	13.0	5.1¼	6.7	Quartiles, 6 st. 9¼
14	14.5	5.2¼	7.8¾	52	13.10	4.11½	7.3¼	lbs., 5 st. 6½ lbs.
15	14.4	5.0	6.0	53	14.8	4.11½	6.9¾	
16	13.6	4.9	5.6	54	13.8	4.5¾	4.9½	Average of quar-
17	14.0	5.2½	7.7½	55	14.8	5.4½	7.0	tiles, 6 st. 1 lb.
18	13.0	4.8¼	5.3	56	14.0	4.10	6.2½	
19	14.7	4.11	6.12¾	57	13.10	4.9	5.5	Half of the ex-
20	14.10	5.1	6.9	58	13.2	5.0½	6.4	amples lie within
21	13.9	4.11	5.11	59	13.6	4.7	5.2½	9 lbs. of median
22	14.10	4.8¾	5.11	60	13.0	4.9	5.9¾	
23	13.4	4.9½	5.8¾	61	13.3	4.8¾	5.5¼	Mode is between
24	13.1	5.2¼	6.1	62	13.5	4.8¼	6.5¾	6 st. and 6¼ st.
25	14.0	4.6¼	5.6½	63	13.10	5.5½	7.10½	
26	14.6	5.3½	7.6½	64	13.1	4.8¾	6.2¼	Average weight
27	14.3	5.0¼	5.11¾	65	13.10	5.4	7.2	between ages 13
28	13.9	4.9	5.11	66	14.0	4.9	5.0½	and 13½ years,
29	13.4	5.1¼	5.9	67	13.3	4.7	5.0	5 st. 9½ lbs.; 13½
30	14.4	5.1	6.8½	68	13.8	4.11	6.1¼	and 14 years, 5 st
31	14.10	4.9½	4.7¼	69	13.7	4.11¾	6.4¼	13½ lbs.; 14 and
32	13.2	4.9½	5.13¼	70	13.11	4.8	4.4½	14½ years, 6 st. 3⅛
33	14.1	4.8¾	5.8½	71	13.11	4.8	4.4½	lbs.; 14½ and 15
34	1?.10	5.2¼	6.8¼	72	13.2	4.7¾	4.10	years, 6 st. 8⅞ lbs.
35	14.0	4.11½	5.7	73	14.0	4.11	6.5	
36	14.4	4.11	6.5	74	13.3	4.3½	4.1¼	Heights may be
37	14.8	4.11	6.0¾	75	13.3	5.0	7.2¾	tabulated in the
38	13.7	5.0¾	6.2	76	13.7	4.8¼	5.6	same way.

Heights arranged in order of magnitude (in.)—

51½, 52½, 53¾, 54¼, 55, 55, 55, 55¾, 55¾, 56.
56, 56¼, 56¼, 56¼, 56½, 56¾, 56¾, 56¾, 56¾;
56¾, 57, 57, 57, 57, 57, 57¼, 57½, 57½, 57½,
58, 58, 58, 58, 59, 59, 59, 59, 59;
59, 59, 59½, 59½, 59½, 59½, 59½, 59½, 59¾, 59¾,
60, 60, 60, 60¼, 60½, 60¾, 61, 61, 61¼;
61¼, 61½, 61½, 62, 62¼, 62¼, 62½, 62½, 62½, 63,
63½, 63¾, 63¾, 64, 64, 64½, 65, 65¼, 65¼.

A graphic method of finding the median of these heights
Graphic method. closely is given by Mr. Galton in the *Report of the Anthropometric Committee* of the British Association, 1881, p. 247; and is illustrated by the diagram facing this page.

On a horizontal line mark off equal intervals representing units of measurement, say inches. On a vertical scale mark off equal intervals representing the number of instances, *i.e.*, persons whose heights are measured. Beginning at the lowest, 51½ in., on a vertical line mark as many dots at equal intervals on the vertical scale as there are persons at that height (in this case only one), so that each dot represents one person. From the highest dot thus marked, suppose a horizontal line drawn till it is over the next height division at which there is an instance, 52½ in., and with this new base proceed as before, marking each instance at 52½ in. by a dot vertically above the 52½-in. mark. Next draw a connected line through the middle points of the consecutive vertical rows of dots; if there is an odd number of dots, the middle one is taken as the middle point; if an even number, the middle point is half-way between the middle ones.

On the vertical scale mark the positions of the median, quartiles, etc., obtained by dividing the distance representing the total number of instances into appropriate parts, and through these points draw horizontal lines to intersect the connected line already drawn. The points of intersection lie vertically above the heights required, as marked on the horizontal scale.

Now it may be assumed that the heights of all persons returned at, say, 58¾ in., are in reality evenly distributed between the limits 58⅝ and 58⅞ in., heights lying within which would be so returned; and it can be verified that the construction just given shows the place of the median, deciles, etc., almost exactly on this hypothesis.

The following analysis is only important when the number of instances is small, and the position of the quartiles, etc., is not evident. There are two cases, (1) where the observations are exact, (2) where the observations are given in grades or to the nearest scale mark.

(1) The following 45 numbers are the numbers of minutes occupied by trains on a certain distance according to time-table :—

45, 46, 47, 48, 48, 51, 53, 54, 55, 58; 61, 61, 62, 65, 65, 69, 69, 69, 71, 76; 76, 76, 77, 77, 78, 80, 81, 81, 82, 82; 83, 83, 84, 85, 85, 85, 85, 87, 88, 89; 90, 92, 94, 101, 103.
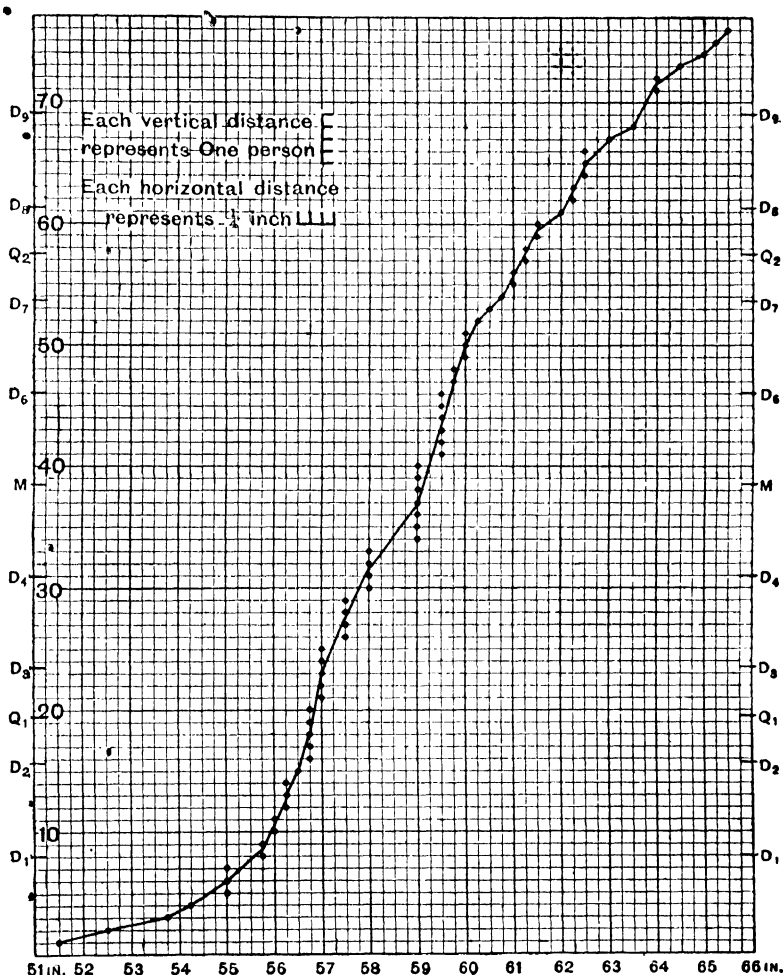
Median 59⅛ inches.

Quartiles 56·8. / 61·2.

Half inter-quartile distance 2·2.

Deciles 55·6, 56·6, 57, 57·9, 63·6, 62, 60·7, 59·7.

Arithmetic average, 59·095.
Greatest density 57 or 59.
,, ,, in smoothed curve would be about 58.
Geometric average 58·98.

The median is the 23rd instance, viz. 77 minutes.

To find the quartiles we must divide the 45 numbers into 4 equal parts. Suppose that on such a scale as the vertical scale in the diagram, p. 106, the instances are entered at $\frac{1}{2}$, $1\frac{1}{2}$, . . . $44\frac{1}{2}$, the distance 45 representing the whole space to be divided. The quartiles are at $11\frac{1}{4}$ and $33\frac{3}{4}$. $11\frac{1}{4}$ is between the 11th instance (61) at $10\frac{1}{2}$ and the 12th instance (also 61) at $11\frac{1}{2}$, and the lower quartile is 61. $33\frac{3}{4}$ comes between the 34th and 35th instances, both 85. If the entries were not equal we might take $\frac{3}{4}$ of the nearer entry (the 34th) $+\frac{1}{4}$ of the 35th.

Similarly the deciles are at the marks $4\frac{1}{2}$, 9, $13\frac{1}{2}$ . . . on the scale. The lowest is at the 5th entry (48 min.), the next half-way between the 9th and 10th ($56\frac{1}{2}$ min.), and so on.

The positions of the D's, Q's and M on the diagram are marked on this principle.

We have the following scheme for the median and quartiles :—

No. of Cases.	Median.	Lower Quartile.	Upper Quartile.
$4n$	$\frac{1}{2}(2n^{\text{th}} + \overline{2n+1}^{\text{th}})$	$\frac{1}{2}(n^{\text{th}} + \overline{n+1}^{\text{th}})$	$\frac{1}{2}(3n^{\text{th}} + \overline{3n+1}^{\text{th}})$
$4n+1$	$\overline{2n+1}^{\text{th}}$	$\frac{3}{4}n^{\text{th}} + \frac{1}{4}\overline{n+1}^{\text{th}}$	$\frac{1}{4}\overline{3n+1}^{\text{th}} + \frac{3}{4}\overline{3n+2}^{\text{nd}}$
$4n+2$	$\frac{1}{2}(\overline{2n+1}^{\text{th}} + \overline{2n+2}^{\text{nd}})$	$\overline{n+1}^{\text{th}}$	$\overline{3n+2}^{\text{nd}}$
$4n+3$	$\overline{2n+2}^{\text{nd}}$	$\frac{1}{4}\overline{n+1}^{\text{th}} + \frac{3}{4}\overline{n+2}^{\text{nd}}$	$\frac{3}{4}\overline{3n+2}^{\text{nd}} + \frac{1}{4}\overline{3n+3}^{\text{rd}}$

A similar scheme could be worked out for the deciles.

(2) If the numbers are given in grades (whether as between, say, 53 and 54 in., or as at 53 in. to the nearest $\frac{1}{2}$ in., i.e., between $52\frac{3}{4}$ and $53\frac{1}{2}$ in.), they may be regarded as spaced uniformly through the grade, and then the method of case (1) applied.

The method can be illustrated from the ages of married men, column 6 of the table, p. 86. Here 549 men are over 40 years, 413 over 45 years; the 500th man is one of the 136 in the grade 40 to 45 years, in fact the 48th or 49th man in that grade. If they are uniformly distributed, the 49th is at the 49th of 136 equal intervals in which the 5 years may be divided. Hence the median is at $40 + \dfrac{549-500}{136}$ of $5 = 41\cdot80$ years. It is not worth while to try to place it more exactly. Similarly the lower quartile is the age where we find the 750th man, somewhere in the grade 30–35 years, and may be taken as $30 + \dfrac{855-750}{152}$ of $5 = 33\cdot45$ years, and the upper quartile at $50 + \dfrac{295-250}{96}$ of $5 = 52\cdot34$.

Simple graphic methods may readily be found for either case.

F. Geometric Mean.

If a_1, a_2 . . . a_n are n quantities G the geometric or logarithmic mean is given by

$$G = \sqrt[n]{a_1 . a_2 . . . a_n},$$

and $\log G = \dfrac{1}{n} (\log a_1 + \log a_2 + . . . + \log a_n).$

The geometric mean is always less than the arithmetic mean of the same quantities.

This mean is appropriately used when emphasis is on the

ratio between two quantities rather than on their absolute difference. , If the difference between 8 and 13 is of the same importance as that between 13 and 18, then the mean of 8 and 18 is properly taken to be 13, equidistant from either; but if the ratio 8 to 12 is of the same importance as that of 12 to 18, then the mean of 8 and 18 is properly taken as $12 = \sqrt{8 \times 18}$.

We obtain an analogy as follows :—Of five quantities a_1, a_2, a_3, a_4, a_5, let a_1 and a_2 be less than A (the arithmetic mean of all) and also less than G, and the others be greater.

Then $\qquad 5A = a_1 + a_2 + a_3 + a_4 + a_5,$

and $(A - a_1) + (A - a_2) = (a_3 - A) + (a_4 - A) + (a_5 - A)$

and $\qquad G^5 = a_1 \times a_2 \times a_3 \times a_4 \times a_5,$

and $\qquad \dfrac{G}{a_1} \times \dfrac{G}{a_2} = \dfrac{a_3}{G} \times \dfrac{a_4}{G} \times \dfrac{a_5}{G}.$

Thus in one case the sum of the excesses of the mean equals the sum of its defects; in the other the product of the ratios of the mean to the quantities less than it equals the product of the ratios of the greater quantities to the mean.

An important use of the mean is in connection with prices. A general rise of prices from 100 to 120 is exactly the same from many points of view as a rise from 120 to 144, and is greater than a rise from 120 to 140. This consideration may have led Jevons to use the geometric mean in his first treatment of index-numbers (*Fall in the Value of Gold*).

It should be noticed that the geometric mean gives greater importance to small numbers and less to large than does the arithmetic.

G. GENERAL.—The function of means will now be clear; it is to express a complex group by a few simple numbers.

The function of   The mind cannot grasp the magnitudes of millions

means.   of items at once; they must be grouped, simplified, averaged. The means chosen must be those which will give the striking features and the essential characteristics of the group. Different methods will apply to groups of various classes; each must be taken on its own merits. A good and suitable mean has the following characteristics :— *If there is a type it shows it; it gives due influence to extreme cases; it is not easily affected by errors or much displaced h·*

*slight alterations in systems of calculation; and it is easily
.calculated.*

The relative positions of the different kinds of means dealt with gives some information as to the general nature of the group to which they refer. The arithmetic mean, median and mode, are coincident, if the group is symmetrical. The arithmetic mean is probably above the median, if we have a small group at a high degree. The arithmetic mean is generally below the median, if there is an absence of high numbers, and a concentration a little above the mean. The mode will be badly defined, if our group is not homogeneous. 'The mode will probably be below the arithmetic mean, if there is a small group at a high degree. The mode is well marked, if the distribution is uniform. These rules are only tentative and easily nullified by exceptional circumstances.

CHAPTER VI.

MEASUREMENTS OF DISPERSION AND OF SKEWNESS. APPLICATION OF AVERAGES.

MEASUREMENTS OF DISPERSION AND OF SKEWNESS.

IN the sections of Chapter V which relate to " means " we have been concerned principally with considering the central position of a statistical group, where by the term *statistical group* we mean a number of persons or things possessing certain defined attributes (Enumerated, in England or Wales, in 1911, male) and grouped according to a variable attribute (age). We can exhibit such a group either by tabulation in grades or otherwise (pp. 69–70) or by a diagram (p. 127), but for purposes of brevity or for comparison with other groups we need to define and calculate measurements related to the group in such a way as to show its characteristics. For this purpose it is convenient to choose (i) a mean which locates a central position, (ii) a measurement of the dispersion, variation or scattering of the observations, and (iii) a measurement of imperfect symmetry. We proceed to the discussion of (ii) and (iii).

Statistical groups.

The differences between the measurements of the items of the group and a mean or other fixed point are called *deviations*. In the table (p. 111) the group taken contains the death-rates of the aggregate of large towns in the 52 weeks of the year 1902. These are arranged in order of magnitude down column 1 and up column 2. In columns 3 and 4 are shown the deviations from the quantity 173, selected as being near the median, 172½. It was shown on p. 104 that the total and therefore the average of deviations (all taken 'positively) is least when they are measured from the median; to obtain such deviations we must add ½ to each entry in column 3 and subtract ½ from each entry in column 4, *i.e.* add and subtract 13 to or from the totals. The total of the positive

Deviations.

deviations from the median is then 447, of negative is 388, and of the 52 deviations (irrespective of their sign) is 835. The average of these deviations, viz. $835 \div 52 = 16 \cdot 06$. To

DEATH-RATES WE.K BY WEEK IN 1902 IN THE AGGREGATE OF GREAT TOWNS IN ENGLAND AND WALES.

Weekly Death-rates per 10,000 living in Order of Magnitude.		For Mean Deviation from Median.		For Standard Deviation.		For Mean Difference.		
Col. 1.	Col. 2	Col. 3.	Col. 4.	Col. 5.	Col. 6.	Col. 7.	Col. 8.	Col. 9.
a.	*b.*	Excess of *a* over 173.	Excess of 173 over *b.*	Squares of Differences from 173.		Difference between *a* and *b.*	Multiplier.	Product.
244	136	71	37	5,041	1,369	108	51	5,508
233	139	60	34	3,600	1,156	94	49	4,606
226	141	53	32	2,809	1,024	85	47	3,995
209	143	36	30	1,296	900	66	45	2,970
206	144	33	29	1,089	841	62	43	2,666
201	145	28	28	784	784	56	41	2,296
196	149	23	24	529	576	47	39	1,833
196	150	23	23	529	529	46	37	1,702
196	151	23	22	529	484	45	35	1,575
191	152	18	21	324	441	39	33	1,287
183	154	10	19	100	361	29	31	899
182	155	9	18	81	324	27	29	783
182	159	9	14	81	196	23	27	621
181	160	8	13	64	169	21	25	525
179	164	6	9	36	81	15	23	345
177	165	4	8	16	64	12	21	252
177	166	4	7	16	49	11	19	209
177	166	4	7	16	49	11	17	187
176	167	3	6	9	36	9	15	135
176	169	3	4	9	16	7	13	91
176	169	3	4	9	16	7	11	77
174	169	1	4	1	16	5	9	45
174	170	1	3	1	9	4	7	28
174	170	1	3	1	9	4	5	20
173	172	0	1	0	1	1	3	3
173	172	0	1	0	1	1	1	1
9,029		434 +13	401 −13	26,491				32,659

835

Arithmetic average $9029 \div 52 = 173 \cdot 63$ approx.
Median 172½.
Quartiles 159½, 181½.
Mean deviation from the median : $\eta = 835 \div 52 = 16 \cdot 06$ approx. ; from the average, $\eta = 16 \cdot 11$.
Quartile deviation, or probable error. $r = \frac{1}{2}(181\frac{1}{2} - 159\frac{1}{2}) = 11$. Half the cases are within 173 ± 11.
Standard deviation: $\sigma = \sqrt{\{\frac{1}{52} \text{ of } 26{,}491 - \cdot 63^2\}} = 22 \cdot 56$. Coefficient of variation $= \frac{100 \, \sigma}{173 \cdot 63} = 13 \cdot 0$.
Mean difference : $g = 32{,}659 \div \frac{1}{2}$ of $52 \times 51 = 24 \cdot 63$.

obtain the sum of the deviations from the arithmetic average ($173 \cdot 63$) we must add ·63 to each of the deviations   Mean deviation.
from 173 of the 28 quantities less than 174 and
subtract ·63 from the remaining deviations; the total is then
$837 \cdot 52$ and the average $16 \cdot 11$. The average of the differences
between the various measurements and their arithmetic average
($16 \cdot 11$ in this case) is called the *mean deviation* of the group;

and often denoted by the letter η; we may also use the term *mean deviation from the median* (16·06). The mean deviation is an obvious and convenient measurement of the dispersion of the group, and where the observations are recorded singly and not merged in grades is easy to calculate.

We can obtain the arithmetic average from columns 3 and 4 at once by the consideration that the average excess over 173 is the total of column 3 (434) less the total of column 4 (401) ÷ 52 = 33 ÷ 52 = ·63 approx.; the average is therefore 173·63 approx.

In the mathematical treatment of statistical groups it is found inconvenient to handle these absolute deviations since in algebra they appear some as positive and others as negative, and when the theory of probability is applied it is found that the importance of the deviation depends on its square and not on its first power. Accordingly the average of the squares of the deviations from the arithmetical average of the group is taken, and the square root of the average obtained is called the

Standard deviation.

standard deviation of the group; this measurement of dispersion is in general use and is denoted by the letter σ. It can be calculated by writing down the deviations exactly, but the procedure is greatly simplified as follows. Let $x_1, x_2 \ldots x_n$ be the measurements, x_0 the central quantity from which the deviations are most conveniently measured. Write $d_1 = x_1 - x_0$, $d_2 = x_2 - x_0$, etc., *i.e.* for the deviations as tabulated. Let $\bar{x}$ be the arithmetic average of the group, so that $n\bar{x} = x_1 + x_2 + \ldots + x_n$; and let $d_0 = \bar{x} - x_0$, so that $nd_0 = (x_1 - x_0) + (x_2 - x_0) + \ldots = d_1 + d_2 + \ldots + d_n$.

Then by definition

$$\sigma^2 = \{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2\} \div n$$
$$= \{(d_1 - d_0)^2 + (d_2 - d_0)^2 + \ldots\} \div n$$
$$= \{d_1^2 + d_2^2 + \ldots - 2d_0(d_1 + d_2 + \ldots) + nd_0^2\} \div n$$
$$= \{d_1^2 + d_2^2 + \ldots - nd_0^2\} \div n, \text{ since } d_1 + d_2 + \ldots = nd_0$$

and $\sigma = \sqrt{\left\{\dfrac{Sd^2}{n} - d_0^2\right\}}$, where . Sd^2 is written for

$$d_1^2 + d_2^2 + \ldots + d_n^2.$$

In the table $\bar{x} = 173\cdot63$, $x_0 = 173$, $d_0 = \cdot63$.

$d_1^2, d_2^2 \ldots$ are given in columns 5 and 6, and . $Sd^2 = 26,491$.

$\therefore \sigma = \sqrt{\{26,491 \div 52 - \cdot63^2\}} = 22\cdot56$ approx.

The standard deviation is always measured in relation to the arithmetical average, not to the median.

A much simpler measurement of dispersion is obtained by the use of the quartiles. The difference between the quartiles is evidently related to the dispersion, though it has the weakness that the same measurement would be obtained from groups whose quartiles were the same, however the observations between the quartiles were distributed, and however far the observations outside the quartiles were placed. It is therefore much less sensitive than the mean or the standard deviations. The measurement used, however, is not the whole distance between the quartiles but half that distance, and   Quartile we may call the half-distance between the quartiles   deviation. the *quartile deviation ;* it is commonly denoted by the letter r. r is approximately, but not in general exactly, equal to the median of the deviations. In the table the quartiles are $159\frac{1}{2}$ and $181\frac{1}{2}$, the distance between them is 22 and $\therefore r = 11$.

The median is not necessarily half-way between the quartiles. In the case before us this half-way mark is $\frac{1}{2}(159\frac{1}{2}+181\frac{1}{2})=170\frac{1}{2}$, and the quartiles are $170\frac{1}{2}\pm11$. We may then describe the group very simply, as follows : the arithmetic average is $173\cdot6$ (or the median is 173) and half the observations are within the range $170\frac{1}{2} \pm 11$.

In a symmetrical group the arithmetic average and the mode are coincident, and r is called the *probable error*, a term that is convenient in some respects, but suggests misleading ideas.

If the data are given in grades a modification of method is necessary and the measurements can only be approximate. Take for example the table of ages on p. 86. The median and quartiles have already been found (p. 107) as   Graded data. $41\cdot80$, $33\cdot45$, and $52\cdot34$ years. The quartile deviation is therefore $\frac{1}{2}(52\cdot34-33\cdot45)=9\cdot45$ years, and half the cases are in the range $42\cdot90\pm9\cdot45$ years. The deviations from $42\frac{1}{2}$ years are given in column 2. If we assume all the entries in each grade to be concentrated at the middle point of the grade, column 4 shows the aggregate deviations in each grade and the sum of the numbers irrespective of sign, viz. $1155 + 926 = 2081$ is the total of the deviations. The mean, deviation from $42\frac{1}{2}$ years is then approximately $2081 \div 1000$ of 5 years $= 10\cdot40$ years. A small correction is needed to obtain the mean deviation from the median or from the average $(13\cdot6\ldots)$. Rather troublesome additions are needed to allow

I*

for the deviations of the 136 entries in the zero grade and to correct for the supposed concentration at the middle points. These become negligible if the grading is sufficiently fine (entries for every year would be sufficient in this case), and it is only then that the use of the mean deviation for graded data is recommended. The origin should be taken at the centre of that grade which contains the average or median, whichever is the starting-point for measuring the deviations.

No new principles are involved in the calculation of the standard deviation in such cases, when the grading is fine. Examples are given on Part II, Chap. I below.

Professor Corrado Gini has introduced a new measurement of variation (*Variabilità e Mutabilità*; Fascicolo 1°; Bologna, 1912, pp. 19 *seq.*). He contends that the problem that arises in the study of the variability of demographic, anthropological, biological or economic characters is How much do the different magnitudes differ between themselves? and not How much do diverse measurements differ from their arithmetic mean? The second question is appropriate in physical science, but not in the description of groups. Accordingly he proposes as a measurement the arithmetic mean of the $\frac{1}{2}n(n-1)$ differences **Mean difference.** that are to be found between n quantities. This we may call the *mean difference* and denote it by the letter g. It has not yet come into general use, possibly because (except in the simplest cases) the arithmetic involved in its calculation is indirect and rather arduous; but it cannot be denied that the conception is simple and logical.

Let $a_1, a_2 \ldots a_n$ be n quantities, arranged in ascending order. Then $g \times \frac{1}{2}n(n-1) =$

$$(a_n - a_1) + (a_n - a_2) + \ldots + (a_n - a_{n-2}) + (a_n - a_{n-1})$$
$$+ (a_{n-1} - a_1) + (a_{n-1} - a_2) + \ldots + (a_{n-1} - a_{n-2})$$
$$\cdot \; \cdot \; \cdot \; \cdot$$
$$\cdot \; \cdot \; \cdot \; \cdot$$
$$+ (a_3 - a_1) + (a_3 - a_2)$$
$$+ (a_2 - a_1)$$
$$= (n-1)a_n + (n-3)a_{n-1} + (n-5)a_{n-2} + \ldots$$
$$+ (1-n)a_1 + (3-n)a_2 + (5-n)a_3 + \ldots$$
$$= (n-1)(a_n - a_1) + (n-3)(a_{n-1} - a_2) + (n-5)(a_{n-2} - a_3) + \ldots$$

The computation is readily performed as in columns 7, 8, 9 of the table, p. 111, where n is an even number. If n is odd the central number occurs by itself with a zero multiplier.

The relation between g (the mean difference) and η (the mean deviation) can be exhibited as follows:—

Let d_1, d_2 . . . d_n be the differences, all taken as positive, between the median and a_1, a_2 . . . a_n.

Then $a_n - a_1 = d_1 + d_n$; $a_{n-1} - a_2 = d_2 + d_{n-1}$, etc.,

and $g = \dfrac{1}{n}$ of $\left\{ 2(d_1 + d_n) + 2 \cdot \dfrac{n-3}{n-1}(d_2 + d_{n-1}) + 2 \cdot \dfrac{n-5}{n-1}(d_3 + d_{n-2}) + \ldots \right\}$,

while $\eta = \dfrac{1}{n}\left\{ d_1 + d_n + d_2 + d_{n-1} + d_3 + d_{n-2} + \right\}$.

In g more than average importance is given to the extreme variations, and g is always greater than η. $E.g.$, if the observations are spaced at equal intervals (k), it can be shown that g is approximately $\eta \times \frac{4}{3}$; for, in this case, if $n = 2m + 1$, it is found that $g = \frac{2}{3}(m+1)k$, $\eta = \dfrac{m(m+1)}{2m+1}k$, $g \div \eta = \frac{4}{3}\left(1 + \dfrac{1}{2m}\right)$; also $g - \eta$ is approximately $\frac{1}{6}mk$.

If the instances are entered, not singly, but as y_1 cases at a_1, y_2 cases at a_2 . . . y_t cases at a_t, where $y_1 + y_2 + \ldots + y_t = N$, the working is more complicated. It can be shown that—

$g \times \frac{1}{2}N(N - 1) =$

$y_t d_t(N - y_t) + y_{t-1}d_{t-1}(N - 2y_t - y_{t-1})$

$+ y_{t-2} d_{t-2}(N - 2y_t - 2y_{t-1} - y_{t-2} + \ldots$

$+ y_1 d_1(N - y_1) + y_2 d_2(N - 2y_1 - y_2) + y_3 d_3(N - 2y_1 - 2y_2 - y_3)$

$+ \ldots$

where the d's in the first and second lines are the differences to quantities above and below the median respectively. The factors are readily computed and arranged in a table.[*]

When measurements are distributed according to the normal curve of error (Part II, Chap. II) we have the following relations:—$\eta = \sigma\sqrt{\dfrac{2}{\pi}}$ $= \sigma \times \cdot 798 \ldots$, $r = \sigma \times \cdot 6745$, $g = \eta\sqrt{2} . = \eta \times 1 \cdot 414 \ldots$ These relations are often obtained approximately in other distributions. Thus on p. 111, $\eta = \cdot 7\sigma$, $g = \eta \times 1 \cdot 41$; but $r = \cdot 5\sigma$ only.

If, following Professor Gini's idea, we take the square root of the average of squares of all the differences, we obtain (whatever the distribution) the quantity $\sigma\sqrt{2\left(\dfrac{n}{n-1}\right)}$, or $\sigma\sqrt{2}$ very nearly.

So far all the measurements of dispersion have been expressed as concrete quantities, as so many shillings, years,

---

[*] The working of the formula here given differs in an unimportant way from that used by Gini, $loc. cit.$, p. 30 and foot-note on p. 29.

points on a scale, etc. It is sometimes advantageous to express them in relation to a mean. Thus if the median and quartiles of a wage group were 30s., 40s. and 50s., the quartile deviation is $\frac{1}{4}$ of the median, while in another group, say 35s., 45s., 55s., it would be $\frac{2}{9}$; it is, in fact, reasonable to regard the second group as being less dispersed than the first, though their quartile deviations are equal. Possible measurements of this class are (a) $\dfrac{\text{Quartile deviation *}}{\text{Mean of quartiles}}$ (b) $\dfrac{\text{Mean deviation}}{\text{Median}}$

(c) $\dfrac{\text{Standard deviation}}{\text{Arithmetic average}}$; but the only measurement at all

Coefficient of variation. generally used is the standard deviation expressed as a percentage of the arithmetic average (*i.e.* (c) $\times$ 100) and this is called the *coefficient of variation*. In the table on p. 111 it is $22 \cdot 56 \times 100 \div 173 \cdot 63 = 13 \cdot 0$.

Asymmetry or *skewness* of a curve is indicated when the mode, median and arithmetic average do not coincide. It is shown more definitely when the sum of the positive deviations from the median is not numerically equal to the sum of the negative deviations; it is also shown when the quartiles, or pairs of deciles, are not equidistant from the median. Any of these inequalities could be made into a measure-

Skewness. ment of skewness. Skewness, relating to the shape, and not to the size, of a curve is appropriately measured by an absolute quantity (resembling the eccentricity of an ellipse), and we therefore need a ratio of two concrete measurements. The simplest to compute is as follows: let q_2 be the excess of the upper quartile over the median, and q_1 the excess of the median over the lower quartile; then $s = \dfrac{q_2 - q_1}{q_2 + q_1}$ is a measure of skewness.† If the curve is symmetrical, $q_2 = q_1$ and $s = 0$; if $q_2 > q_1$, s is positive, and if $q_2 < q_1$, s is negative. s becomes $+ 1$, if $q_1 = 0$, that is if the median and lower quartile coincide, and s becomes $- 1$, if $q_2 = 0$. s is therefore a measurement which never exceeds 1 numerically, and has a definite significance at zero and at its extreme values. In the table on p. 111, $q_2 = 9$, $q_1 = 13$, $s = \dfrac{-4}{22} = - \cdot 19$. In the

---

* In earlier editions I called this quantity, the *dispersion*. It has the advantage that it is necessarily not greater than 1.

†

table of ages (see pp. 86 and 107) $q_2 = 10·54$, $q_1 = 8·35$, $s = ·12$. The significance of various values can only be obtained by experience, but it may be suggested that ·1 is a moderate degree of skewness, and ·3 a considerable degree.

It should be noticed that the three characteristics of a group can be measured simply from the quartiles and median; the median for the central position, the quartile deviation for the dispersion, and the measurement just discussed for the skewness.

Some Examples of the Application of Averages.

If our analysis of the nature and use of averages is complete and if averages are of widely extended use, we *Application of* should now be able to express almost any group *averages* of figures by a few well-chosen numbers of definite significance.

To apply a somewhat severe test at first, let us choose a familiar example from ordinary life, and consider how a suburban business man might test the merits of *to train service.* two railway systems, by one of which he intended to take a season ticket.

The following table gives the train service between Leatherhead and London in 1898 :—

TRAIN SERVICE—LEATHERHEAD TO LONDON.
Number of Minutes to Journey.

WATERLOO —
  Down—60, 50, 52, 48, 47, 61, 50, 44, 48, 53, 45, 42, 45, 49, 43, 48, 42, 43.
      Sundays—50, 50, 47, 49, 50.
  Up—51, 46, 51, 48, 43, 44, 48, 48, 64, 45, 48, 47, 45, 47, 46, 47.
      Sundays—48, 48, 51, 51, 51.
LONDON BRIDGE—
  Down—67, 65, 65, 61, 74, 51, 56, 66, 65, 53, 59, 41, 49, 44, 58, 57, 56, 67, 80.
      Sundays—67, 52, 66, 68, 88, 65, 65, 68, 65.
  Up—69, 57, 53, 58, 54, 41, 58, 52, 42, 40, 55, 67, 79, 98, 69, 66, 68, 64, 71.
      Sundays—72, 71, 69, 70, 62, 81, 73, 73.
VICTORIA—
  Down—77, 65, 55, 76, 77, 88, 48, 53, 46, 69, 89, 54, 82, 71, 90.
      Sundays—92, 45, 81, 84, 78, 61, 85, 83, 85.
  Up—87, 65, 69, 69, 47, 48, 51, 83, 101, 58, 62, 61, 76, 103.
      Sundays—81, 76, 80, 85, 85, 82, 94.

The following table gives us the necessary information :—

	London Bridge.	Victoria.	Waterloo.
	Min.	Min.	Min.
Average of four quickest trains -	41	46½	42½
Lower decile · · · ·	47½	48	43
Median · · · ·	65	77	48
Mode · · · ·	65	...	48
Number of trains on week days-	38	29	34
General average · · ·	63	73	48

It is to be noticed that the statistical method is generally limited to one aspect of a problem; the question of punctuality might, indeed, be easily treated statistically, but the questions of comfort and relative picturesqueness of route will elude our analysis.

The next example shows a method of throwing into relief the characteristics of a typical group of sociological data.

Tabulation of wages returns. The adjoining table gives the wages recognised by the Amalgamated Society of Engineers in many of their branches in 1862 and 1891.

AMALGAMATED SOCIETY OF ENGINEERS.—WAGES in 1862 and 1891, Weekly, exclusive of Overtime.

	1862. s. d.	1891. s. d.		1862. s. d.	1891. s. d.
Accrington	27 0	31 0	Faversham	34 0	33 0
Ashford	33 6	30 0	Folkestone	34 0	32 0
Ashton-under-Lyne	29 3	34 0	Frome	24 0	{ 27 0 / 30 0 }
Bacup	26 1	28 0			
Barrow-in-Furness	31 0	34 9	Gainsborough	27 6	28 0
Bath	29 0	31 0	Glossop	27 2	32 0
Bedford	27 0	29 0	Gloucester	28 0	•32 0
Bilston	28 0	30 0	Grantham	28 6	30 4
Bingley	24 0	29 0	Grimsby	28 0	32 0
Birkenhead	29 0	35 6	Halifax	23 1	31 0
Birmingham	32 0	36 0	Hanley	28 3	32 0
Blackburn	27 6	32 0	Hartlepool	26 0	34 10
Bolton	27 6	{ 28 0 / 32 0 }	Heywood	27 0	{ 30 0 / 34 0 }
Bridgwater	24 6	24 0	Holyhead	32 0	28 0
Brighton	24 8½	29 0	Huddersfield	26 0	26 0
Bristol	31 0	32 0	Hull	27 •6	34 0
Burnley	27 0	30 0	Hyde	{ 30 0 / 28 0 }	28 0
Burton-on-Trent	25 0	30 0			
Bury	28 3	{ 30 0 / 32 0 }	Ipswich	28 6	28 0
Cardiff	31 0	34 0	Keighley	23 0	27 0
Carlisle	24 6	30 0	Kidderminster	28 0	30 0
Chepstow	30 0	34 0	Lancaster	25 0	32 0
Chester	30 0	32 0	Leeds	25 0	30 0
Chowbent	26 0	32 0	Leicester	26 0	31 6
Colne	25 0	31 0	Leigh	27 9	31 6
Congleton	24 0	28 0	Lincoln	26 7	28 6
Coventry	28 0	34 0	Liverpool	29 0	34 0
Crewe	29 4	30 0	Llanelly	22 0	26 0
Darlington	25 0	31 6	Macclesfield	24 0	29 6
Dartford	34 0	38 0	Manchester	29 9	35 0
Darwen	27 0	32 0	Mexborough	27 0	32 0
Derby	26 0	29 0	Middlesborough	25 0	34 0
Doncaster	28 6	31 6	Middleton	29 5	33 0
Dover	35 6	36 0	Milton and Elsecar	28 0	34 0
Enfield Lock	36 0	40 6	Neath	32 0	30 0
			Newark	25 0	29 0
Exeter	23 0	{ 28 0 / 32 0 }	Newcastle	25 0	{ 35 0 / 37 0 }

AMALGAMATED SOCIETY OF ENGINEERS,—WAGES in 1862 and 1891, Weekly, exclusive of Overtime (*continued*).

	1862. s. d.	1891. s. d.		1862. s. d.	1891. s. d.
New Holland - -	30 8	34 0	Stafford - - -	34 0	30 0
Newport - - -	30 0	32 0	Stalybridge - -	28 3	32 0
New Town (Stockport)	29 0	32 0	Stockport - - -	28 0	{32 0, 34 0}
Newton Abbott - -	33 0	33 0			
Northampton - -	26 0	32 0	Stockton-on-Tees -	24 0	36 0
Northfleet - - -	36 0	36 0	Stoke-on-Trent -	29 0	32 0
North and So. Shields	26 0	35 0	Stroud and Thrupp -	26 0	30 0
Norwich - - -	32 0	29 0	Swindon - - -	31 6	31 6
Nottingham - -	27 5	34 0	Todmorden - -	26 0	28 0
Oldbury - - -	28 0	34 0	Wakefield - -	25 0	30 0
Oldham - - -	29 0	33 0	Warrington - -	28 0	34 0
Peterborough - -	28 6	33 0	Watford - - -	35 0	36 0
Plymouth - - -	32 0	33 0	Wednesbury -	26 0	31 0
Pontypridd - -	24 0	30 0	Whitehaven - -	25 0	{28 0, 36 0}
Portsmouth - -	35 0	34 0			
Preston - - -	27 0	32 0	Wigan - - -	28 0	34 0
Radcliffe Bridge -	27 0	{30 0, 32 0}	Wolverhampton -	28 0	33 0
			Wolverton - -	29 2	29 0
Reading - - -	28 0	{32 0, 34 0}	Worcester - -	31 0	30 0
			Bermondsey -	35 4	
Ripley - - -	26 0	26 6	Blackwall - -	34 0	
Rotherham - -	27 6	32 0	Bow - - -	36 0	
Rugby - - -	32 0	{28 0, 32 0}	Greenwich - -	34 0	
			King's Cross -	36 0	
Rugeley - - -	24 11	30 0	Lambeth - -	35 8	
St Helens - -	28 0	{34 0, 36 0}	London, E. - -	35 0	38 0
			,, N. - -	35 10	
Sheffield - - -	28 0	36 0	,, S. - -	35 0	
Shipley - - -	25 9	{28 0, 30 0}	,, W. - -	35 6	
			Marylebone -	33 0	
Shrewsbury - -	30 6	32 0	Stratford - -	{35 0, 33 6}	
Smethwick - -	28 0	35 0			
Southampton - -	32 0	34 6	Tower Hamlets -	36 6	
Sowerby Bridge -	24 6	30 0	Woolwich - -	36 0	

The following figures show the same in brief :—

	1. 1862.* s. d.	2. 1891.* s. d.	3. 1891.† s. d.
Maximum - - - -	36 6	40 6	...
Upper decile - - -	35 0	38 0	38 0
Upper quartile - - -	31 4	34 0	36 0
Median - - - -	28 0	32 0	34 3
Arithmetic average - - -	28 10	32 4	33 4
Modes - - - -	28 0	{30 0, 32 0}	...
Lower quartile - - -	26 0	30 0	31 6
Lower decile - - -	24 6	28 6	30 0
Minimum - - - -	22 0	24 0	...
Quartile deviation - - -	2 8	2 0	2 3
Skewness, from quartiles - -	.25	0	— .22

* Each branch counting as 1.
† The numbers of members in each branch counted as receiving the wage recognised there.

If the rates at each branch were not those actually paid to all members, but their average, while the actual wages were confined within small limits of that average, the figures in the last column would be little affected.

On comparing columns 1 and 2 it will be seen that not only have all the averages increased, but that since the lower decile and quartile have increased more rapidly than the upper, the lower half has also gained on the upper. Again the wages are grouped more closely in column 2 than in column 1.

GROUP C OF TABULATION.—It was necessary to postpone the tabulation of non-numerical or descriptive answers till we had finished our discussion of averages. The following detailed example shows how the median, etc., can be used to give a short description of a large group of adjectival answers.

Tabulation of descriptive answers.

In 1891 the Amalgamated Society of Engineers obtained from all their branches answers to the question : To what extent is overtime worked? The branch secretaries sent answers which may be tabulated as on next page.

An inspection of the table here given will show sufficiently the method of tabulation. The position of most of the answers in an imaginary scale is fairly definite, except that it is not always obvious where the numerical answers should be placed; this must be decided either by internal evidence or practical knowledge of the trade. The same adjectives did not of course convey exactly the same numerical meaning to all the branch secretaries who used them, but it will be admitted that this tabulation gives a fairly clear view of the case, and that the method of medians and quartiles may be appropriately applied   Taking the member of a branch as the unit and neglecting the unclassed answers, the median is " Maximum 18 hours in 4 weeks " or " moderately," the lower quartile " Very little," and the upper quartile " 14 hours when busy." Taking the branch as unit, the median is " Not much," the quartiles are " Very little " and " When necessary " or " Occasionally."

Explanation of table.

This method, which, with varying degrees of precision, is widely applicable, seems to afford the only way of comparing two such groups of answers. The precision attainable is to be measured by the distance through which the median can be shifted by making reasonable variations in the scheme of

Answers.	Number of Branches.	Number of Members.
None -	4	140
Not worked	1	78
Very little	23	4,836
To very limited extent	1	63
Very occasionally	1	350
A little on repairs	1	500
Little -	2	73
2 hours when necessary	1	80
Seldom	1	59
Small extent -	1	16
Seldom except on repairs	1	66
Only on repairs	2	216
Not much	6	1,125
On repairs	1	500
Not to any extent	3	644
Not to a great extent -	2	162
Not general -	1	7
Not systematically	2	43
In cases of breakdown or emergency -	7	606
2 hours regularly	1	136
Chiefly on repairs	1	20
Occasionally -	2	90
When necessary	1	348
Casually (sic) -	2	142
A good deal on repairs	1	23
Maximum 18 hours in 4 weeks	1	1,000
Moderately	3	262
Systematically in good trade -	1	200
Average about 5 hours a week	1	96
Considerably in marine shops	1	400
Systematically in dockyard	1	650
General	2	146
Systematically	1	693
Great amount -	1	263
To a great extent	1	72
Excessively	1	550
9 hours a week	1	39
10 ,,	1	106
12 ,, (maximum)	1	700
14 ,, (when busy)	1	106
10 to 18 hours a week -	1	5,000
Total -	88	20,666

UNCLASSED :—

No answers -	36	5,114
As little as possible	1	250
Not so much lately -	1	160
In machine shops for six months	1	60

Now that we have the method of averages at our disposal
Summarisation. we may use it for tabulating and summarising a
group of figures.

Consider, for example, the answers to the questions issued
by the Commissioners on Trade Depression in 1886.

Four of the questions were :—

    1. Number of men in Society.
    2. Number out of work in 1885.
    3. Weekly wage in 1885.
    4. Change in wages between 1865 and 1885.

The following table shows the answers given by the branch
secretaries of the Amalgamated Society of Engineers :—

1. District.	2. No. in District, 1885.	3. No. Out of Work, 1885.	4. Current Wages, 1885.	5. Wage change between 1865 and 1885.
Belfast	1,100	130	28/ to 36/	Slight increase.
Coventry	2,500	230	31/6	Contract work—50 °/₀ decrease.
Dukinfield	170+	20+	31/	Slight increase.
Dundee	1,400	45°/₀	25/ skilled. 15/ unskilled.	Time work—1865, 22/; '72, 24/; '80, 26/; '83, 24/; '85, 25/.
Glasgow	28,000	4,000	26/	Time wages, 5 °/₀ above 1864.
Glasgow (St Rollox)	1,600	250	—	Rise in 1872-73 of 15 °/₀; 1885 same as 1865.
Hartlepool	1,200	400	31/6	Advance of 3/.
Glossop	135	10	32/	....
Liverpool	280	38	...	Rise in 1872-73 of 7½ °/₀; 1885 same as 1865.
Monifieth	114	18	21/	Skilled work—1865, 24/; '76, 27/; '78, 25/; '83, 28/; '85, 25/.
Nottingham	4,000	600	34/ minimum.	1865, 28/; 1885, 34/.
Oldham	1,600	96	33/ average.	Increase of 5 °/₀.
Oxford	45	—	33/	....
Paisley	800	...	28/6	1865, 26/; 1885, 28/6.
Preston	630	40	28/	None.
Preston	900	120	28/	None.
Shipley	201	15	28/6 24/ non-unionists.	1865, 28/6; 1869-73, 32/; 1885, 28/6.
Sowerby Bridge	1,120	43	28/	1865-75, 25/6; 1875-85, 28/.
Sunderland	3,200	400	33/	1864, 27/; '74, 34/; 1875-85, between 31/ and 37/.
Swindon	6,050	2	31/6	....
Ulverston	45	...	31/	1865, 26/; 1875, 31/.
Wednesbury	400	30	30/	Increase of 2/.
Workington	170	70	28 to 36/	Increase of 30 °/₀.

It is suggested that the following are the summary tables which should be inserted in a report dealing with the answers.

The figures are given here for only one society, but the tabulations are framed so as to include all.

TABLE I.—STATE OF EMPLOYMENT.

Name of Society.	Total Number * in Branches making Returns on Employment.	Number Out of Work.	Percentage Out of Work.	Median of the Percentages Out of Work in the Various Branches.
A.S.E. -	55,170	7,142	13	12
O.S.B. -				
&c. -				

* Details of some of the most important branches should be added.

TABLE II.—CURRENT WAGES.

Name of Society.	Average of Wages in Branches.		Quartiles of Branch Wages.	Measure of Dispersion. (v. p. 116 (a)).
	Unweighted.	Weighted.		
A.S.E. -	s. d. 30 0	s. d. 29 7	s. d. s. d. 28 0 32 0	1⁄14
O.S.B. -				
&c. -				

TABLE III.
A. CHANGE OF WAGE BETWEEN 1865 AND 1885.

Name of Society.	Number of Branches showing				Median of Percentage Increases.	Percentages of Members in Branches showing			
	No Answer.	De-crease.	No Change.	Increase.		No Answer.	De-crease.	No Change.	Increase.
A.S.E.	4	1	5	13	10	11	4	6	79
O.S.B.									
&c.									

Verbal Summary.—In the great majority of cases a considerable increase of wage took place between 1865 and 1885,

equivalent on the whole to a rise of about 10 per cent. The figures are not sufficiently definite to give an exact average.

TABLE III.—*B.* CHANGE OF WAGE BETWEEN 1865 AND THE MAXIMUM ABOUT 1873.

TABLE III.—*C.* CHANGE OF WAGE BETWEEN MAXIMUM ABOUT 1873 AND 1885.
(Tabulation as in III. *A.*)

CHAPTER VII.

THE GRAPHIC METHOD.

I. General Purpose.

The two main methods of elementary statistics which ought to be understood by all students or officials who handle figures, which are easily within the grasp of all independently of mathematical training, but are generally misunderstood or ignored by the uninterested or the uninitiated, are the method of averages and the method of diagrams or the graphic method. These two are placed together because the uses of averages and diagrams are nearly related. When we deal with large and complex masses of figures we are unable to grasp them in their entirety, however clearly they may be tabulated. Any list of figures—the populations of different towns, the death-rates at successive ages, the wages of many workpeople, the imports for a series of years—becomes less comprehensible as its length increases. A series of ten numbers can, perhaps, be easily grasped, of twenty only with an effort ; while a printed list of figures for one hundred successive years leaves hardly any impression on our mind at all ; we cannot see the wood for the trees. The test to which all questions as to the use of averages should be referred is that the averages selected should afford the best summary of the whole group in question that the mind can grasp. When the meaning of the word average was sufficiently extended, we found that we could select three, four, or even ten suitable figures which adequately showed the main features of any group. The main use of diagrams is also to present large groups of figures so that they shall be intelligible in their entirety, and the test for all diagrams is that the diagram as drawn should afford the best view of the series or group of figures that the eye can appreciate. Diagrams have one use which averages have not, for it is only by a diagram that a series of figures relating to successive years can be adequately

Averages and diagrams.

presented ; but in reality they are less essential than averages, for the latter often have an existence independently of the figures from which they are derived, representing true types of the quantities which are being measured; and by their use alone are further comparisons of complex groups made possible : while diagrams, on the other hand, might be dispensed with, being auxiliary rather than essential, merely an aid to the eye and a means of saving time.

To connect this chapter more closely with the preceding, we will show how the same group of figures, for example the wages of a large group of workpeople, may be represented by either method. '

Graphic representation and averages.

Consider the following data :—

Numbers of workpeople earning—

From 15/ to 16/	-	200 ⎫		From 25/ to 26/	-	1,200 ⎫	
,, 16/ ,, 17/	-	400 ⎪		,, 26/ ,, 27/	-	800 ⎪	
,, 17/ ,, 18/	-	100 ⎬ 1,000		,, 27/ ,, 28/	-	700 ⎬ 3,500	
,, 18/ ,, 19/	-	100 ⎪		,, 28/ ,, 29/	-	500 ⎪	
,, 19/ ,, 20/	-	200 ⎭		,, 29/ ,, 30/	-	300 ⎭	
,, 20/ ,, 21/	-	200 ⎫		,, 30/ ,, 31/	-	300 ⎫	
,, 21/ ,, 22/	-	300 ⎪		,, 31/ ,, 32/	-	400 ⎪	
,, 22/ ,, 23/	-	300 ⎬ 2,200		,, 32/ ,, 33/	-	400 ⎬ 2,100	
,, 23/ ,, 24/	-	500 ⎪		,, 33/ ,, 34/	-	500 ⎪	
,, 24/ ,, 25/	-	900 ⎭		,, 34/ ,, 35/	-	500 ⎭	

From 35/ to 36/	-	600 ⎫
,, 36/ ,, 37/	-	400 ⎪
,, 37/ ,, 38/	-	100 ⎬ 1,200
,, 38/ ,, 39/	-	80 ⎪
,, 39/ ,, 40/	-	20 ⎭

Using the method of averages we should replace this group by the following figures :—

								s.	*d.*
Average of all	-	-	-	-	-	-	-	27	6
,,	lowest 1,000	-	-	-	-			17	0
,,	highest 1,000	-	-	-	-			36	6
,,	middle 4,000	-	-	-	-			27	0

or

Median, 26/9; quartiles, 24/2, 32/.
Deciles, 20/, 23/6, 24/9, 25/8, 26/9, 28/2, 31/, 33/4, 35/4.
Mode, 25/3; secondary positions, 16/6, 36/.

or

Persons earning from	15/ to 20/	20/ to 25/	25/ to 30/	30/ to 35/	35/ to 40
Percentages of all -	10	22	35	21.	12

This group is represented on the annexed diagram, an example of the graphic representation of the relation between two variable quantities. A figure similar to this Construction of simple diagrams. may be used to show marriage, or death-rates at different ages, numbers of persons of various statures, demand at different prices, or any such group of homogeneous quantities. The same construction can be used to show the changing values of any number in a series of years. Draw a line parallel to the bottom of the page, and mark equal intervals to represent a quantity which can have many successive small increments, such as age, income, height, price, time, and so on. This is called the axis of *abscissæ*, and the distance of a point measured from the zero position along the line is called its *abscissa*. At right angles to this line, parallel to the side of the paper, through the zero position we draw another, called the axis of *ordinates*, and grade this to correspond to the numbers possessing the qualities represented by the abscissæ; at each grade on the axis of abscissæ; draw lines at right angles to it, to represent on the chosen scale the numbers at that grade; these lines are called the *ordinates*. In the annexed diagram the abscissæ represent the amounts of wages, the ordinates the number of persons earning them. Join the tops of the ordinates by straight lines and the diagram is complete. In practice, when squared paper is used, without drawing the ordinates their tops can be marked.

This diagram shows at one glance the distribution of the wage-earners according to their wages. A small number earned between 15s. and 16s., a slightly larger group Description of the wage diagram. between 16s. and 17s., very few between 17s. and 19s. Above 19s. the number continually rises; high numbers are found from 24s. to 27s., the highest between 25s. and 26s. The line falls to the 30s. group, but not so low as between 17s. and 19s., then it rises regularly to 36s., and falls rapidly to 39s. Here, then, we have the main group congregated in the neighbourhood of 25s., a distinct but smaller group at 36s., and a small and nearly isolated group at 16s.; representing a considerable group of highly-skilled men between 30s. and 40s., the great mass with ordinary skill between 20s. and 30s., and a small group of incompetents at 16s. These features would not be so easily seen from the tabulated figures.

It is to be noticed that the number tabulated as between 15s. and 16s. is represented by the ordinate at 15s. 6d., the middle of the interval; if the original figures on which the table was based had been given to the nearest 1d., the ordinate should be drawn at 15s. 5½d. It is important that these middle points should be accurately placed.

The use of the line joining the tops of the ordinates is two-fold. First, it enables the eye to judge relative heights more easily; and secondly, it suggests the idea of continuity, which can be better illustrated by the next diagram. In this the abscissæ represent ages, the ordinates the estimated numbers of persons living at and above the ages at which they stand per thousand inhabitants of England and Wales at the middle of the year 1891. The ordinates were drawn at the points on the axis of abscissæ representing the middle of each year of age; but length of life cannot be expressed exactly in years, or even in months, days, or minutes. The intention of the diagram is to show the proportion living above each age, and for this purpose the joining line should have no breaks or sharp angles, but should suggest absolute continuity.

Continuity.

In practice, it is useless to mark in the points for smaller intervals than a year, for the eye could not grasp the detail. It is, however, implied that the line drawn has the same shape as that which would result if the number of persons was infinite and the subdivision by age infinitesimal.

Estimated number per 1,000 of the population at and above—

Ages.		Ages.		Ages.		Ages.		Ages.	
0	1,000	16	628	32	346	49	152	65	47
1	973	17	607	33	332	50	143	66	43
2	949	18	587	34	318	51	135	67	38
3	925	19	567	35	305	52	127	68	34
4	901	20	547	36	292	53	119	69	31
5	877	21	528	37	280	54	112	70	27
6	854	22	510	38	268	55	104	71	24
7	830	23	491	39	256	56	98	72	21
8	807	24	474	40	244	57	91	73	18
9	783	25	456	41	233	58	85	74	15
10	760	26	439	42	222	59	79	75	13
11	738	27	423	43	211	60	73	76	11
12	715	28	407	44	201	61	67	77	9
13	693	29	391	45	191	62	62	78	8
14	671	30	376	46	181	63	57	79	6
15	649	31	361	47	171	64	52	80	5
				48	161				

Calculated from the Census of 1891.

NUMBERS PER 1,000 OF THE POPULATION ABOVE ASSIGNED AGES.
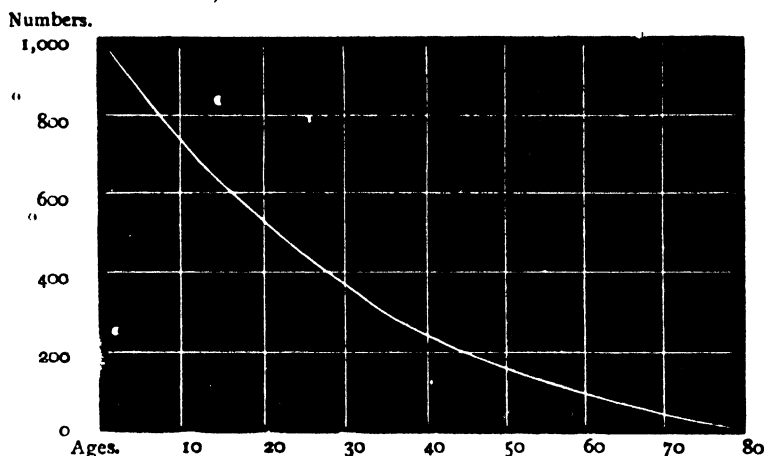
Apply these remarks to the diagram facing p. 127. Average earnings for a year will not be reckoned exactly by shillings or even pence; if we had a sufficient number of instances we should get regular sequences of earners at successive farthings, and the line representing them would have no sharp angles, but be continually curved. The figure rightly gives the eye this impression of continuousness. Similarly in the diagram representing exports facing p. 134, the line correctly gives the impression that exports are continuous day by day.

By an obvious step we may suppose that the unit of *area*, that contained between vertical lines through two consecutive divisions on the axis of abscissa, and horizontal lines through two consecutive divisions on the axis of ordinates, represents one wage-earner, and it is then easy to see that the area contained between the base line, the curve, and two vertical lines through the points marking any two amounts of wage represents the total number earning rates between those amounts.

Hence the lines (diagram, p. 127) through M, the position of the median, Q_1, Q_3 those of the quartiles, D_1, D_2, D_3, D_4, M, D_6, D_7, D_8, D_9 of the deciles divide the area $ABm_1m_2m_3CD$ into two, four, and ten equal areas respectively. The centre of gravity of this figure lies on the vertical line through V, the average wage; and the feet of the ordinates through the highest points m_1, m_2, m_3 are at the modes.

K*

When the grades in which the data are tabulated are wide it is better to use the method of the next diagram, which we may call a *block diagram.*

This and the drawing underneath it illustrate the numbers of married men distributed by age which are given on p. 86.

Graded data.   In that table we have no information except that such a proportion are as old as twenty years and not as old as twenty-five years, etc. This is precisely represented by constructing a rectangle with base the interval that represents five years, and height proportional to the number recorded within that interval. The method of the diagram facing p. 127 would suggest that all were at the middle of the grade. In the case of ages we know that the succession of numbers year by year ought to be continuous, and a complete representation would be a continuous curve, such that the area standing on a five years' interval equals the area of the corresponding rectangle. Such a curve is drawn free-hand on the diagram. If the figure is such as to leave little margin of uncertainty as to the position of the curve throughout, then the curve is an adequate representation of the facts.

The data may also be represented by the lower diagram, where the crosses show the information as recorded in the table. These crosses are joined by straight lines; the resulting figure may, if the phenomena are continuous, be replaced by a curve, which in this case would hardly be distinguishable from the straight lines.

The details of technique of diagram drawing, the position of the scales, the devices for making the figure clear, and so Requisite   on, can be gathered from the various diagrams accuracy.   given in this chapter. The degree of accuracy to which the figures should be marked, whether correct to a million, a thousand, or a unit, is determined simply by the power of the eye to grasp detail; in most of those here given it will be found that a displacement of one in a thousand is perceptible, and this is the ordinary limit. More minute accuracy is useless, for it is not the function of diagrams to dispense with lists of numbers, but only to enable the eye to perceive their significant features.

Before discussing the choice of scales on which the numbers are to be represented, it is necessary to consider the ways in

DISTRIBUTION BY AGE OF MARRIED MEN, ENGLAND AND WALES 1911.

I Block Diagram. Number per 1 000 in 5-Year Grades.
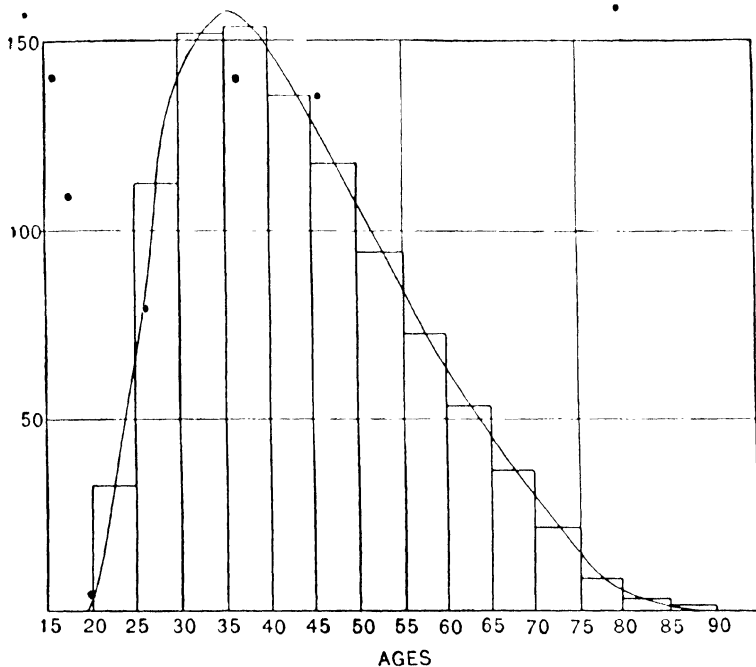
AGES

II Cumulative Diagram Numbers per 1,000 above the Age shown.

AGES
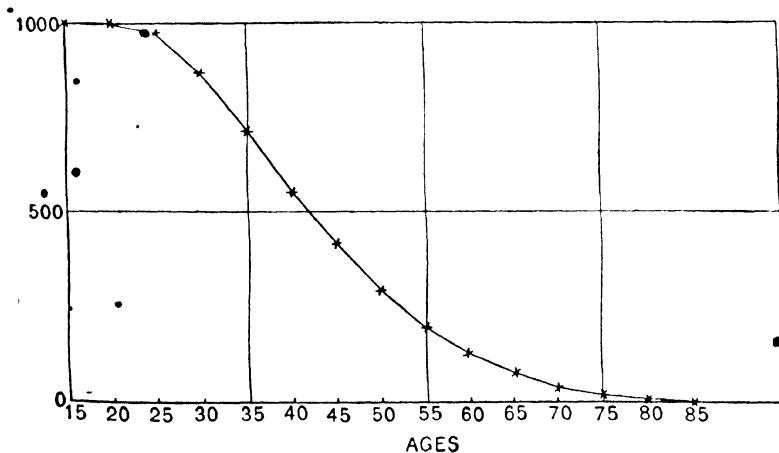
which a diagram makes an impression on the eye. The eye
·can judge—(1) Distances; (2) ratios; (3) angles. Optical power.
The dotted lines in the diagram facing p. 134 will
illustrate these points. (1) The eye is a fairly safe judge of
distances; there is very little doubt which of two points is the
further from the base line; when squared paper is used, a
difference of 1 in 1000 is perceptible. The eye can also judge
differences quickly. In the figure the value of the exports in
1883 exceeded that in 1885 by more than the value in 1890
exceeded that in 1883. (2) It can be seen that the value of
exports approximately doubled between 1862 and 1889; or
that the value in 1878 is about three-quarters of that in 1890.
The accuracy with which the eye can make such measurements
is not great; it is not easy to detect that the ratio of the values
in 1873 and 1871 (1·095 : 1) is greater than the ratio of the
values in 1882 and 1880 (1·073 : 1); but the general impression
given by the diagram is partly made up by unconscious calcula-
tions of this nature. To make these observations accurately
the method described on pp. 169 *seq.* should be used. Notice
that for these observations the insertion of the base line is
necessary; and, because they are made unconsciously, a dia-
gram showing movements over a series of years without a base
line gives an incorrect impression. (3) The question, Was the
increment greater in 1886–87 or in 1887–88? can be more
quickly answered by observing the angles than by noting the
differences. The line showing the latter change is steeper
(makes a greater angle with the horizontal) than the line showing
the former. Hence the latter increase is the greater; actually
£12,600,000 against £9,200,000. The most useful exercise
of this power, however, is to judge the dates at which the rate
of increase changed; thus the value of exports increased in
1862–63, increased at a slower rate in 1863–64, and slower
yet in 1864–65, more rapidly in 1865–66; a slow fall followed
in 1866–67, then an increase began which is continually
accelerated to 1871, and so on. The line from 1872–76 is
concave to the base line, showing an accelerated fall; the
concavity from 1879 to 1882 corresponds to a retarded
rise. The increases so shown are absolute or actual, not
relative or in ratio to the quantities at the beginning of each
period.

It is difficult to lay down rules for the proper choice of the

scales by which the figure should be plotted out.   It is only the
Choice of • ratio between the horizontal and vertical scales'
scale.   that need be considered.   The figure must be
sufficiently small for the whole of it to be visible at once ; if the
figure is complicated, relating to a long series of years and
varying numbers, minute accuracy must be sacrificed to this
consideration.   Supposing the horizontal scale decided, the
vertical scale must be chosen so that the part of the line which
shows the greatest rate of increase is well inclined to the vertical,
which can be managed by making the scale sufficiently small ;
and, on the other hand, all important fluctuations must be
clearly visible, for which the scale may need to be increased.
Any scale which satisfies both these conditions will fulfil its
purpose.   The page opposite shows the erroneous impressions
which can be given by a judicious manipulation of the scale
and by the omission of the base line.   The diagrams, which
are drawn roughly, all represent the same estimates of wages in
England and in the United States of America for certain years
from 1860.   Figure 1 sets the lines in proper relief.   In Figure 2,
Necessity of   the base line is not drawn in the zero position
correct   for the English scale, and the American scale is
base line.   reduced ; the consequence is that English wages
appear to have fluctuated widely, while American made steady
progress.   In Figures 3, 4, and 5 the scales are doctored and the
base line adjusted, so that in 3 American wages seem to have
caught up English, in 5 exactly the reverse is the case, while in
4 wages appear to have moved with equal rapidity in both
countries.   An examination of these figures will show that the
eye cannot be trusted to supply the right base line, or to
estimate the importance of fluctuations without it ; and, with
certain exceptions to be mentioned later,* it is well to distrust
all those numerous diagrams, where space has been economised
at the expense of the base line.
     We can now pass on to the consideration of the smooth-
ing of curves, for which purpose the question of the " alleged
Smoothing   stationariness of our exports," discussed by Sir R.
curves.   Giffen in his paper before the Royal Statistical
Society in 1899, affords an excellent illustration.   The thin
dotted line on the diagram opposite shows the value of exports

---

·THE SAME FIGURES REPRESENTED ON VARIOUS SCALES AND WITH ERRONEOUS BASE 'LINES.

In each figure the scale for English wages is on the left, that for American wages on the right.

English Wages.
American Wages.

Figure 1. Figure 2. Figure 3. Figure 4. Figure 5.
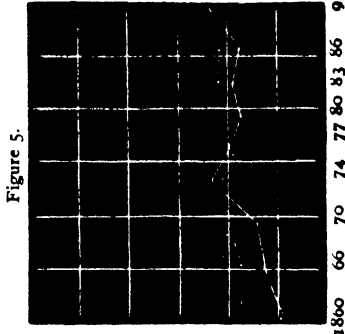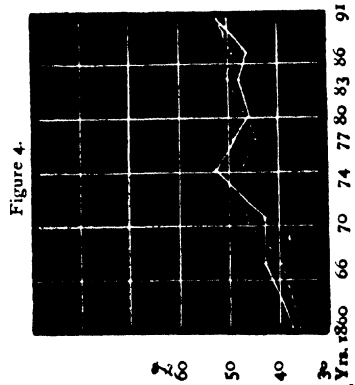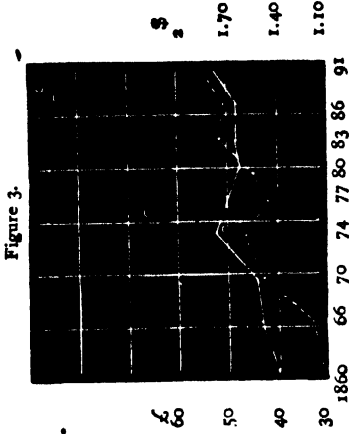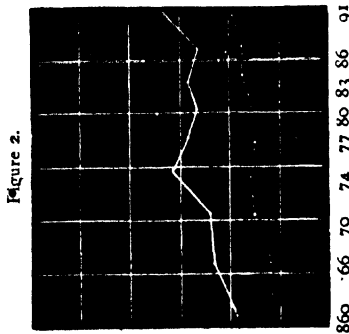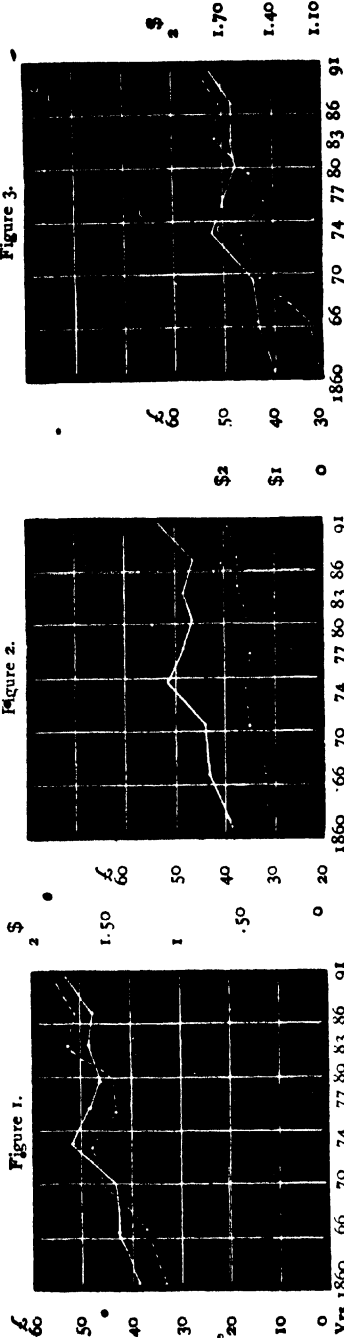
AVERAGE WAGES		
Years	In England per Annum.	In U.S.A. per Diem.
1860	£38.8	$1.10
1866	43.2	1.22
1870	43.9	1.50
1874	52.4	1.60
1877	49.8	1.48
1880	47.6	1.57
1883	48.8	1.75
1886	48.0	1.71
1891	53.2	1.85

TOTAL DECLARED REAL VALUE OF BRITISH AND IRISH PRODUCE EXPORTED FROM THE UNITED KINGDOM.   I = £1,000,000.

		Averages.					Averages.		
		Three Yearly.	Five Yearly.	Ten Yearly.			Three Yearly.	Five Yearly.	Ten Yearly.
1855	95.7	...	...	...	1881	234.0	216.2	208.2	221.6
1856	115.8	...	...	...	1882	241.5	232.9	216.7	220.1
1857	122.0	111.2	...	...	1883	239.8	238.4	226.0	218.6
1858	116.6	118.1	...	...	1884	233.0	238.1	234.3	217.9
1859	130.4	123.0	116.1	...	1885	213.1	228.6	232.3	216.9
1860	135.9	127.6	124.1	...	1886	212.7	219.6	228.0	218.1
1861	125.1	130.5	126.0	...	1887	221.9	215.6	224.1	220.4
1862	124.0	128.3	126.4	...	1888	234.5	223.0	223.0	224.5
1863	146.5	131.9	132.4	...	1889	248.9	235.1	226.2	230.2
1864	160.4	143.7	138.4	127.2	1890	263.5	249.0	236.3	234.2
1865	165.8	157.6	144.4	134.3	1891	247.2	253.2	243.2	235.5
1866	188.9	171.7	157.2	141.6	1892	227.1	245.9	244.2	234.1
1867	181.0	178.6	168.7	147.5	1893	218.1	230.8	240.9	231.9
1868	179.7	183.2	175.1	153.8	1894	215.8	220.3	234.3	230.2
1869	190.0	183.6	181.0	159.8	1895	225.9	219.9	226.8	231.4
1870	199.6	189.8	187.8	165.9	1896	240.1	227.3	225.4	234.1
1871	223.1	204.2	194.6	175.7	1897	234.3	233.4	226.8	235.4
1872	256.3	226.3	209.7	188.9	1898	233.4	235.9	229.8	235.3
1873	255.2	244.9	224.8	200.0	1899	255.3*	241.0	237.8	236.1
1874	239.6	250.4	234.7	207.9	1900	283.6*	257.4	249.3	238.1
1875	223.5	239.4	239.6	213.7	1901	270.9*	269.9	255.5	240.5
1876	200.6	221.0	235.1	214.9	1902	277.7*	277.4	264.2	245.5
1877	198.9	207.7	223.7	216.7	1903	286.5*	278.4	274.8	252.3
1878	192.8	197.4	210.9	218.0	1904	296.3*	286.8	283.0	260.4
1879	191.5	194.4	201.4	218.1	1905	324.4*	302.4	291.2	270.2
1880	223.1	202.5	201.3	220.5	1906	367.0*	389.2	310.4	282.9

* Not including the value of ships exported.

year by year, and the first impression given by it is that exports have not grown in value in recent years. Sir Robert Giffen gave the following table :—

AVERAGE ANNUAL VALUE OF EXPORTS.

1855–57	-	-	-	-	-	£134,000,000,
1865–67	-	-	-	-	-	228,000,000
1875–77	-	-	-	-	-	264,000,000
1885–87	-	-	-	-	-	274,000,000
1895–97	-	-	-	-	-	292,000,000

and from this he deduced " that all through there is an increase,' and that the only sign of stationariness is an increase at a less rate in the last periods than in the earlier periods."

The *Saturday Review* * wrote " that such a conclusion is grossly misleading," for the figures are merely triennial averages of selected years showing a happy coincidence; " why was not 1898 included? " An inspection of the numbers does not show us the answer to this criticism, but on the diagram the whole

---

circumstances are visible at a glance. Since 1865 three great waves have been completed. The maximum of 1872, due to the inflated prices of that year, is very high, but that of 1890 is greater than any previous figure, while the maximum in 1882 is comparatively low. The minima increase throughout; those of 1868, 1879, 1886 show a regular progression, which falls off greatly in 1891. In 1894–96 it looked as if another decennial cycle was in progress, but this was checked in 1897. Since the discussion, the returns for the successive years to 1906 have shown an increase, surpassing that which preceded 1872.

The *Saturday Review* went on to ask why Sir Robert Giffen did not give " proper quinquennial averages," such as—

<div align="center">

AVERAGE ANNUAL VALUE OF EXPORTS.

</div>

1870–74	-	-	-	-	-	£235,000,000
1880–84	-	-	-	-	-	234,000,000
1890–94	-	-	-	-	-	234,000,000
1898	-	-	-	-	-	233,000,000

and it must be granted that this gives an appearance diametrically opposite to that of the previous table.

It is clear that we need some general method of bringing these figures into a form which shall be quite independent of the choice of any special years. The diagram facing page 134 does this. The thin continuous line, lying almost over the dotted line of annual values, shows triennial averages taken yearly, that is the average of each year with those before and after it ; this line smooths off the corners without affecting the general appearance. The line of crosses shows quinquennial averages, each year being averaged with the two previous and two subsequent years. The line of circles shows decennial averages ; each circle is placed at the centre of the period whose average it represents ; thus the circle showing the average of the ten years 1875–84 is placed vertically over the line separating the years 1879 and 1880.*

. On looking at the line of quinquennial averages it is clear that the *Saturday Review* did precisely what it accused Sir Robert Giffen of doing, for years are taken which favour the argument. The quinquennial periods selected for comparison with 1898 are all on the upper parts

Choice of periods.

* In all the curves of averages the mark showing the average is placed at the centre of gravity of the marks showing the 3, 5, or 10 quantities averaged.

of the waves, the marks showing these averages are very near
the maxima of the quinquennial line, while the year 1898 does
not appear to be a maximum. We might with just as much or
as little accuracy give the following :—

QUINQUENNIAL AVERAGES OF THE VALUES OF EXPORTS.

1865-69	-	-	-	-	-	£181,000,000
1875-79	-	-	-	-	-	201,000,000
1885-89	-	-	-	-	-	226,000,000
1898	-	-	-	-	-	233,000,000

and say that the value in 1898 was higher than any of the pre-
vious selected averages. There is no need to use arbitrary dates
to get at the facts. No argument can stand which does not take
account of the cycle of trade, which is not eliminated till we
take decennial averages. Special marks in the diagram show
the averages for decennial periods, indicating a rapid increase
before 1870, followed by steady slow progress till the subsequent
expansion. The complete line gives just the same general
appearance. If, finally, the figures were completely smoothed
by a freehand line keeping as close to this as was possible,
without making sudden changes of curvature, the same appear-
ance would be given; the thick line on the diagram is an
attempt to do this. The smoothing is obtained by the assump-
tion that the cycle of trade is ten years; when two maxima fall
within the same ten years the average of this period by our
construction gives the appearance of a maximum (*e.g.*, in 1887)
at a date of a minimum. This would be avoided if we con-
tinually changed our period for averaging to accommodate
the changing wave-length, a somewhat arbitrary proceeding.
The difficulty thus arising can be easily corrected by the eye,
and the final smoothed line is intended to convey this corrected
impression.

It should be clear now that it was in 1899 five years too
soon to pay attention to the particular figure for 1898; the
figures for the next five years, necessary to determine the char-
acter of the coming wave, could not be foretold. When these
are included it is seen that each decennial average (for 1890–99,
1891–1900, etc.) established a new record, and that the figures
for each year from 1900 to 1906 are greater than those of any
previous maximum. It will be seen, moreover, that the sentence
quoted from Sir Robert Giffen on p. 134 is fully justified.

The smoothed line now constructed represents the general tendency of the value of exports, when accidental and temporary variations are removed. If it were possible to separate entirely variations of short period from secular changes, to separate the ebb and flow of the

Meaning of smooth line. "Trend."

tide of commerce from the steady current of increasing trade, we may suppose that we should obtain a result represented by this line. In it there are no sudden changes even in rates of growth, while the addition and subtraction year by year of relatively small quantities would produce precisely that irregular .fluctuating line from which the smooth line was obtained.

The diagram can be continued from the following numbers :—

1907	-	-	-	416·0*	369·1	338·0	301·1
1908	-		-	366·5*	383·2	354·0	314·4
1909	-		-	372·3*	384·9	369·2	326·1
1910	-		-	421·6*	386·8	388·7	339·9
1911	-		-	448·5*	414·1	405·0	356·7
1912	-		-	480·2*	450·1	417·8	377·9
1913	-		-	514·2*	481·0	427·4	400·7

* Not including the value of ships exported.

The records during the war are not comparable with those here given. The reader is recommended to study the diagram as printed and to judge how far forecasts of amount, fluctuation and general movement are possible, before looking at the actual records of 1907-13.

The direction of the smooth line at any date may be called the *trend* of the series at that date. When the smooth line is approximately straight over several years, its general direction shows the trend in that period.

A special method of determining the trend has been recently used by Professor Moore (*Statistical Journal*, 1919, p. 375). He assumes that the general movement over a stretch of years can be represented by the equation $y = a + bx + cx^2 + dx^3$, and determines the values of a, b, c and d by the condition that, if y_i is the observed value at a date x_i, then $S(y_i - a - bx_i - cx_i^2 - dx_i^3)^2$ should be a minimum. Professor Persons (*Review of Economic Statistics*, Harvard, Preliminary Volume, No. 1) assumes that a straight line is sufficiently accurate and minimises $S(y_i - a - bx_i)^2$. It is doubtful whether either of these methods is of general application, and Persons' hypothesis in particular must be used with discretion. The method of moving averages (used in the test above) is certainly more sensitive for showing changes in the direction of the

trend if a long series of years is under consideration, and the general causes which determine the phenomena have definitely varied several times.

The fuller discussion of " smoothing " series of figures belongs to the chapter on interpolation, but one other group may
Smoothing a homogeneous group.
here be considered, as showing the use of the graphic method for obtaining regularity out of irregular raw material.   Referring back to the figures given on p. 69, we can exhibit the wages of 5000 workers anew by a diagram, in which the ordinates represent the numbers earning *at or above* a certain wage. ·The thin angular line on the adjacent page represents these numbers, entered for every 10-cent group.   This plan is especially useful for irregular figures, like this wage-group, for the line must always tend upwards from the numbers earning the highest wage to the numbers earning at least the lowest.   The diagram is also at once adaptable to the graphic method of finding the median described on p. 106.

The irregularities shown by the thin line do not arise from any law of wage-grouping, but are due to the accidents of observation ; if we regard these returns as samples out of a much larger unregistered group, we may suppose that a smoothed curve will indicate approximately the form which would be obtained, if our returns were complete.   To smooth this figure, draw a freehand line passing as near the points as possible without abrupt changes of curvature, as in the annexed diagram. A new approximation may be made for the median, quartiles,
Graphic method of finding the median.
etc., by drawing horizontal lines through the points on the vertical scale corresponding to half, one-quarter, three-quarters, etc., of the workers ; from the points where these cross the smooth line, draw vertical lines to the scale of dollars ; the points on the scale so obtained are the median (quartile, etc.) wage.

The results obtained are :—

	Median.	Quartile.	Quartile.
Given on p. 70 - -	$1.49	...	...
By method of p. 106, used in annexed diagram -	$1.49	$1.16	$2.12
From smooth curve in annexed diagram - -	$1.51	$1.15	$2.13
By method of interpolation, p. 227 - - -	$1.536	...	...

GRAPHIC METHODS OF DETERMINING THE MEDIAN AND MODES.
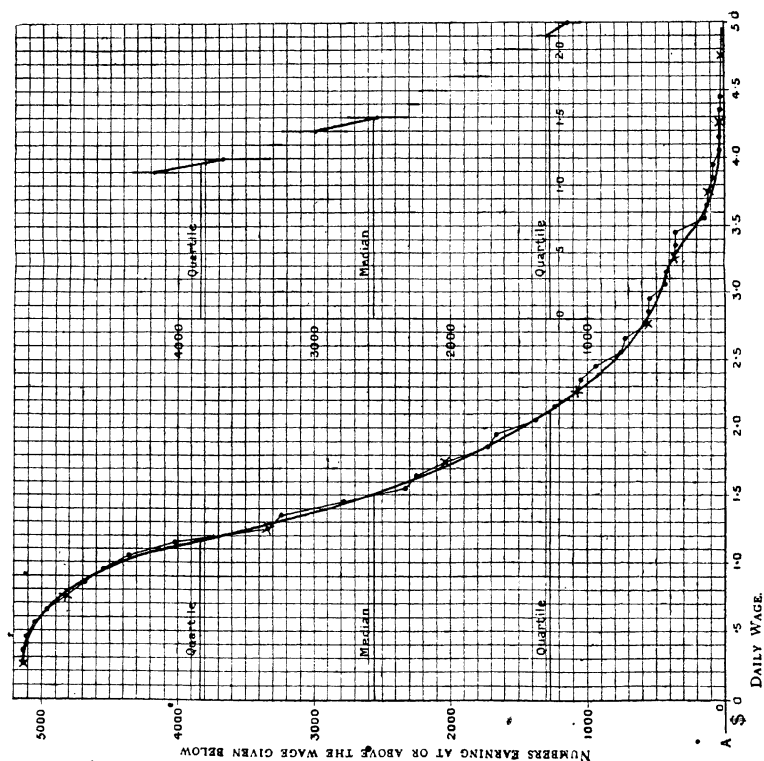
This method is not, however, one of great precision; a very slight change in the curvature of the smoothed line would make more difference than those shown between the second and third lines in the above table.

This method is useful for determining the mode approximately. It will be remembered that the difficulties in doing this before arose from the uneven distribution on the two sides of the mode, and in the displacement of the mode by the adoption of a second system of tabulation.

The first of these difficulties entirely disappears in the graphic method, while the second is diminished, for the displacement now only depends on the slight possible variations in the curvature of the smooth line. The mode is clearly the position where the greatest number is added, in the present method of representing the figures : that is, the mode is where the line, angular or smooth, is steepest. On the smooth curve the maximum steepness is where the tangent crosses the curve,—in mathematical language, at a point of inflexion. This can be determined mechanically by placing a ruler to touch the curve, and turning it round the curve till it crosses it. On the annexed figure this occurs in the interval between $1.10 to $1.40. A more complex method of determining both mode and median, is discussed in Chap. X, pp. 227–8.

This graphic way of finding these means has two great advantages. It can be applied to numbers which are given at irregular intervals of graduation (e.g., 30 at 30s. 6d., 40 at 30s. 8½d., 35 at 40s. 1d., etc.) as easily and by exactly the same construction as to more regular returns; and if the smooth curve is carefully drawn, the *number* of modes can be seen at a glance and the individual importance of each can be estimated. In the annexed diagram, the curve is concave to the base line from $.30 to about $1.20, convex from about $1.20 to $3.15, concave till $3.40, and then convex till the end. The points of inflexion or the modes are where concavity gives way to convexity. Hence there are two modes, of which that near $3.4 is of the less importance.

A large class of diagrams may be passed by with a few words. Writers and lecturers frequently use points, lines, triangles, squares, circles, even pictures, of different sizes to assist the presentation of the relative magnitude of numbers. These have their use for

popular lectures and hand-books, but do not add anything to the significance of the figures. Collections of these may be found in the second volume of Gabaglio's *Teoria Generale della Statistica,* and in M. Levasseur's *La Statistique Graphique* in the *Jubilee Volume* of the Royal Statistical Society.

Of these one group may be signalled as of practical use. Rectangles may be used to express three quantities : one side to represent price ; the adjacent side, quantity ; and the area, value : or number of houses, average number of inmates and population : or number of hours' work per week, average output or hourly wage, and total output or weekly wage. The figures on the annexed page show the limit to which this method can be usefully pushed.

REPRESENTATION OF THREE FACTS BY RECTANGLES.

Imaginary budgets of an artisan and a labourer, showing amounts spent weekly on various commodities, and number of hours' work necessary for each amount.

Per week,
£1. 13s. 4d.
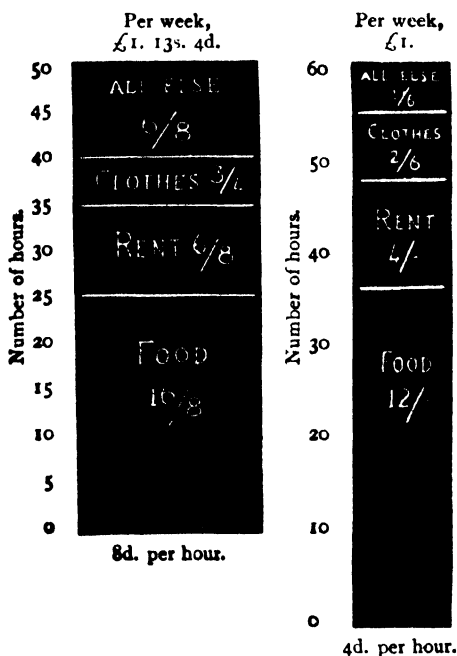
Per week,
£1.

The horizontal scale represents pence per hour. .125 inch = 1d.

The vertical scale represents number of hours per week. .1 inch = 2 hours.

The areas represent amounts spent, and the whole rectangles show the week's wages on the same scale. 1 sq. in. = 13s. 4d. '

A Joint Committee on Standards for Graphic Representation has since 1916 worked at the best methods for presenting

statistics graphically, and has made many useful suggestions which may produce uniformity in treatment and avoid errors.

• The use of statistical maps can only be afforded a brief notice here. Any numerical quality of a population, its density, average income, average taxation, may be shown district by district by suitable markings, $\hspace{1em}$ Cartograms. or colours. Of these the most useful method is to choose one colour, say blue, for excess above the average ; another, say red, for defect. Divide the districts in nine groups, say more than 7 per cent., 5 to 7 per cent., 3 to 5 per cent., 1 to 3 per cent. above the average : these should be marked by four shades of blue, becoming lighter as the average is approached ; within 1 per cent. of the average, above or below, should be white ; and shades of red, gradually becoming darker, will show the remaining grades below the average. Care must be taken not to adopt too many grades. For examples of this method see Booth's *Life and Labour of the People,* maps ; the *Statistical Atlas of the XIth Census* of the United States ; the *Statistical Atlas of India ;* and the maps in M. Levasseur's paper just mentioned. A cheap and very effective method, by which similar results are obtained in black and white only, may be seen on Plate P (misprinted 2) in that paper, and in the excellent chapter on Graphic Representation in Bertillon's *Cours élémentaire de Statistique,* p. 133 *seq.*

A common defect in maps of this class arises from the fact that records generally relate to administrative areas, while the phenomena to be represented are independent of these. An example will make this difficulty evident. If a map is made of England in 1911 colouring the counties according to the density of population, Cumberland will be marked by the colour appropriate to 27 persons per 100 acres, and Northumberland by that for 53 persons. The colour will change abruptly in a moorland region where for many miles the population is of a uniform sparseness. This difficulty can be overcome by either of two methods. Minute divisions, *e.g.,* civil parishes, can be taken as the units, and each shaded in black only, the amount of pigment increasing with the population ; or the population can be marked *in situ* as accurately as the data allow, a dot of uniform size being placed for each 100 people, with a modification of method for dense districts. A map of

this kind is reproduced in Professor Secrist's *Statistical Methods*, 1917, p. 189.

2. HISTORICAL DIAGRAMS.

Gross   Revenue   divided   by   Population

Zero   line   for   Revenue   per   head

Revenue   and   Customs

Gross   Revenue

Total   Inland   United   Kingdom

Population   of   the

Excise

Customs

Income Tax

Post   and   Telegraph

£3·58

£ million   150 − 2

140 − 1

130 — 0

Per head scale

Scale for Population

Years 1850   55   60   65   70   75   80   85   90   95   1900   05
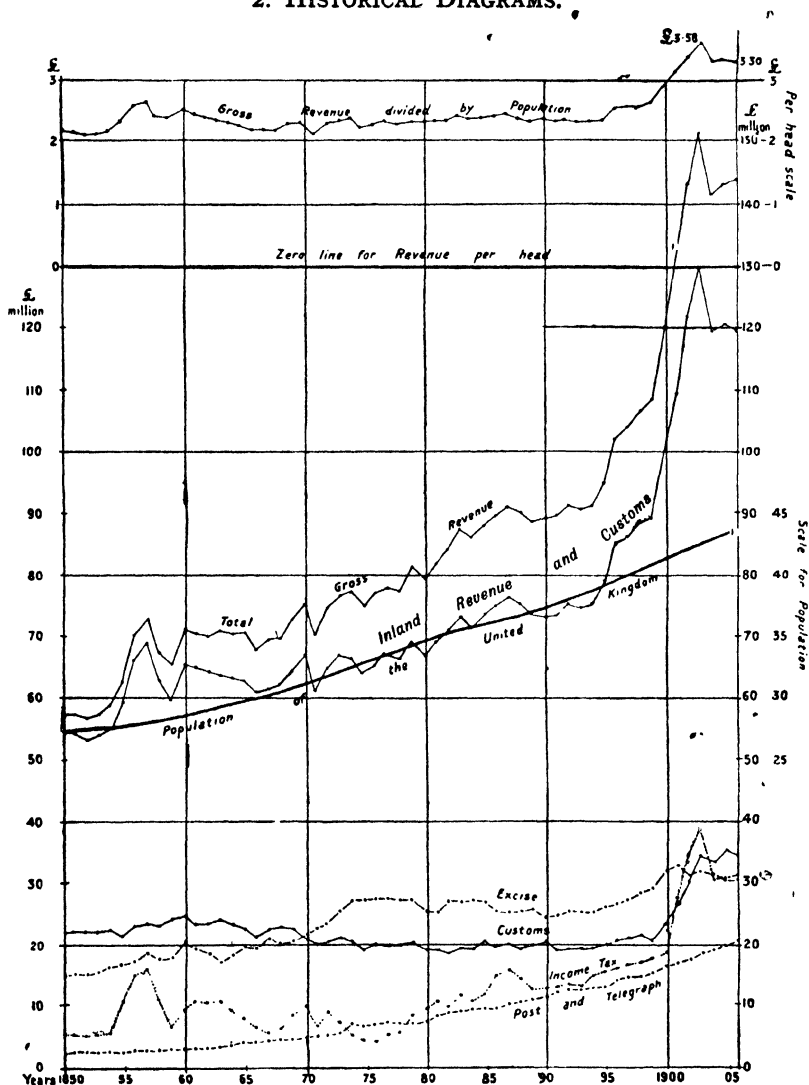
Perhaps the chief use of diagrams is to afford a rapid view of the relations between two series of events.

The different cases that occur are best illustrated by

examples. The simplest is when we wish to compare two sets of figures expressed in the same unit, say £ sterling; _{Comparison of} and the simplest of these when we wish simply to _{figures} compare a whole and its parts.

On the adjacent diagram the upper line shows the annual total gross revenue of the United Kingdom (*Statistical Abstract*, 1906 *); the next line, that part which comes _{illustrated by} from inland revenue and customs, the difference _{the revenue.} being mainly composed of post office receipts. The principal heads of revenue are customs, excise, income tax, and post office. These are shown by suitable lines for each year, each line being independent of the other, and all having the same base line and being on the same scale. This method is greatly preferable to the alternative one of drawing a second line representing the total less customs, a third the total less customs and excise, and so on, because the eye is then quite incapable of judging the relative movements of the separate items. The figure shows at once the main features of the course of revenue. The increase has been rapid but irregular. The rapid growth in 1854–57 was not at once maintained, but the figures for the 60's are at a far higher level than those for the 50's. A rapid fluctuation in 1870 is followed by a more regular growth almost unchecked till 1887; and then, after a short stationary period, there are great increases in 1895, and between 1898 and 1903. Nearly the same remarks apply to the line showing inland revenue and customs. If we look for the parts of the revenue that have borne the increase and change, we see that prior to 1900 receipts from excise had increased most, next those from the post office, and next those from the income tax, while the customs had diminished. Each line has its distinctive features. The post office payments show an almost regular growth. The income tax fluctuates violently, bearing the brunt of nearly all the rapid changes in the total, especially in 1856 and 1870, and 1900–02. The excise line shows a moderate increase till 1870, a sudden jump to 1874, and a very slow growth since that date. Customs, on the other hand, have to some extent taken an opposite course to that of excise, so that the total from the two had not changed very rapidly prior to 1900. At the top of the page a new base line is taken,

---

* This diagram cannot be carried later owing to a change in the book-keeping of Impe_ial and Local taxation accounts.

REVENUE OF THE UNITED KINGDOM.

Unit, in all columns, £10,000.

Year ended 31st March	Total Revenue.	Inland Revenue and Customs.	Customs.	Excise.	Property and Income Tax.	Post and Telegraph.
1850	5,739	5,431	2,226	1,497	560*	216
1851	5,732	5,412	2,204	1,528	560*	228
1852	5,658	5,335	2,222	1,538	550*	237
1853	5,753	5,401	2,214	1,575	570*	237
1854	5,890	5,502	2,251	1,630	580*	252
1855	6,282	5,944	2,163	1,680*	1,070*	237
1856	7,026	6,601	2,324	1,730*	1,520*	281
1857	7,279	6,848	2,353	1,840*	1,620*	292
1858	6,788	6,309	2,311	1,782	1,159	292
1859	6,548	5,987	2,412	1,790	668	320
1860	7,109	6,570	2,446	2,036	960	331
1861	7,028	6,514	2,331	1,943	1,092	340
1862	6,986	6,412	2,367	1,833	1,036	351
1863	7,060	6,390	2,403	1,715	1,057	365
1864	7,021	6,306	2,323	1,821	908	381
1865	7,031	6,291	2,257	1,956	796	410
1866	6,781	6,036	2,128	1,979	639	425
1867	6,943	6,156	2,230	2,067	570	447
1868	6,960	6,204	2,265	2,016	618	463
1869	7,259	6,422	2,242	2,046	862	466
1870	7,543	6,708	2,153	2,176	1,004	477
1871	6,994	6,106	2,019	2,279	635	527
1872	7,471	6,484	2,033	2,333	908	543
1873	7,661	6,660	2,103	2,578	750	583
1874	7,734	6,608	2,034	2,717	569	700
1875	7,492	6,397	1,929	2,739	431	679
1876	7,713	6,525	2,002	2,763	411	719
1877	7,857	6,636	1,992	2,774	528	730
1878	7,774	6,610	1,997	2,746	582	746
1879	8,115	6,899	2,032	2,740	871	757
1880	7,934	6,695	1,933	2,530	923	777
1881	8,187	6,895	1,918	2,530	1,065	830
1882	8,396	7,058	1,929	2,724	994	863
1883	8,739	7,313	1,966	2,693	1,190	901
1884	8,616	7,187	1,970	2,695	1,072	947
1885	8,799	7,380	2,032	2,660	1,200	966
1886	8,958	7,493	1,983	2,546	1,516	989
1887	9,077	7,611	2,015	2,525	1,590	1,028
1888	8,980	7,566	1,963	2,562	1,444	1,060
1889	8,847	7,360	2,007	2,560	1,270	1,118
1890	8,930	7,341	2,042	2,416	1,277	1,177
1891	8,949	7,358	1,948	2,479	1,325	1,226
1892	9,099	7,534	1,974	2,561	1,381	1,263
1893	9,040	7,480	1,971	2,536	1,347	1,288
1894	9,113	7,543	1,971	2,520	1,520	1,301
1895	9,468	7,865	2,011	2,605	1,560	1,334
1896	10,197	8,512	2,076	2,680	1,610	1,422
1897	10,395	8,597	2,125	2,746	1,665	1,477
1898	10,661	8,855	2,180	2,830	1,725	1,518
1899	10,834	8,945	2,085	2,920	1,800	1,586
1900	11,984	9,963	2,380	3,210	1,875	1,665
1901	13,038	10,956	2,626	3,310	2,692	1,725
1902	14,300	12,189	3,099	3,160	3,480	1,779
1903	15,155	12,993	3,443	3,210	3,880	1,838
1904	14,155	11,935	3,385	3,155	3,080	1,915
1905	14,337	12,053	3,573	3,075	3,125	1,993
1906	14,398	11,987	3,447	3,023	3,135	2,101

* These figures cannot be given accurately within £100,000.

and the number of pounds per head of the population is shown year by year; it will be seen that the only important increases were between 1853 and 1857, and from 1898 to 1903.

So far we have found no more difficulty in the choice of scales than previously when dealing with only one line, for all the lines on the larger diagram indicate millions   Choice of of pounds, and when the unit is £1, a new base   second scale. line has been adopted. But we may wish to show the change of population on the larger diagram. It is necessary, as we have already seen, to use the same base line for the two quantities to be compared; but we may choose any point for the beginning of the new line, adapting our vertical scale, for the eye can judge the proportionate changes wherever the line is placed. It is best to decide this point by defining the problem on which the comparison should throw light. If it is required to compare the growth of revenue with the growth of population since, say, 1850, we should start the new line at the point on the 1850 line where the revenue curve begins, and we can then see how the lines intersect one another again and again. Since 1850, however, is an arbitrary date, this plan lacks definition, and it is more logical to make the lines coincide at the most recent date given, with which any previous date can then be compared. On the diagram the line is drawn on such a scale that it lies fairly close to that for inland revenue throughout the greater part of its course.

The next diagram, facing p. 146, introduces further diffi- culties as to the choice of scales. The object of the figure is to show the relations between quantity, value, and   Comparison of price of imported wheat, and population. The line   quantity and A is first drawn on a scale chosen so as to throw its   value. fluctuations into relief. Population is at once brought into relation with this by calculating the amount per head year by year. The line C to represent these figures is drawn on a different scale, chosen so that the line shall not cause confusion by continually crossing any of the others on the figure. If the figure was too full this could be treated as on p. 142, the revenue per head. The same scale of years must be used, and for simplicity of calculation and appearance, 100 lbs. consumed per head is measured by the same vertical distance as   Details of 10,000,000 cwt. imported. A and C refer to the   construction. same quantities, and therefore similar lines are used in both

L*

cases.   The line B represents value and is shown by a broken line.   For this line the choice of scale is more difficult.   In the diagrams which follow, instances will be shown where special methods are used to bring out specific comparisons.   Here this is not necessary, and a scale is adopted which brings the lines A and B into near relation, and shows the fluctuations of B, while the figure is made simple and intelligible by the representation of £20 by the same vertical distance as 20 cwt.

The line D shows the changing price of wheat as deduced from columns A and B.   The scale is chosen so that it boldly crosses the lines A and B; thus its fluctuations are clearly shown, and the numbers are easily seen, for 2s. per cwt. is represented by the same vertical line as 10,000,000 cwt.   If the figure was accurately drawn, lines A and D would lie one over the other in 1876–77; they are therefore shifted very slightly horizontally, and clearness is preserved without the general impression being vitiated.

The lines in the diagram, elucidated by the table, suggest many characteristics and changes which call for explanation Movements needing explanation. by students of economic history.   The consumption of imported wheat per head increased for thirty years to 1895, and was then lower for some years.   The quantity imported shows violent short-period fluctuations.   The price after violent fluctuations from 1862 to about 1878 fell for seventeen years with little intermission. Here no doubt is shown the effect of many causes : an increasing population, the fact that wheat imported is complementary to the home product which is dominated by the English weather, the variation of harvests all over the world, political events, the fall in the value of silver, the development of communication and transport, etc.   The function of the diagram is to show the general trends and the dates of change, but of course one cannot from it ascertain the causes.

As regards the choice of markings for different lines, the chief rule is that lines which cross one another, unless very acutely, must be marked differently.   The second rule is to mark similar quantities in similar ways.

If it is possible to use more than one colour this principle can be easily carried out.*

---

* See *Wages in the Nineteenth Century*, by the present author, diagram facing p. 90.

IMPORTATION OF WHEAT AND WHEAT FLOUR, 1862–1906.

x--x--x Price: shillings per cwt. ___ D.D.
·--·--·-- Value: millions of £ _____ B.B.
_____ Quantity: millions of cwt. ___ A.A.
_____ Quantity: lbs. per head _____ C.C.

To face page 146.
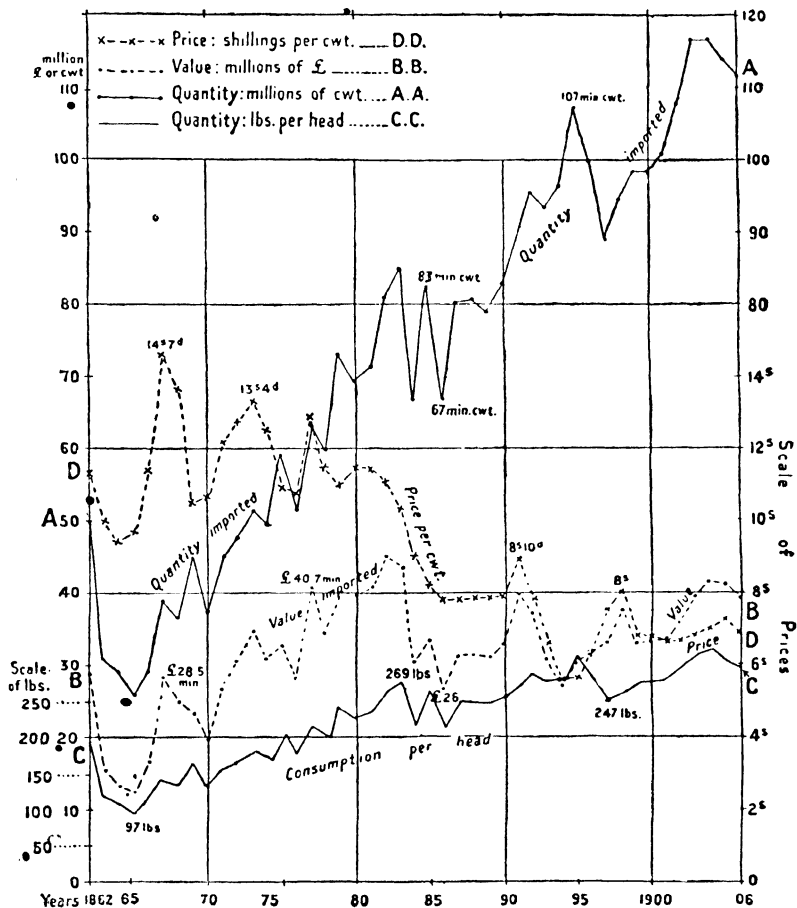
IMPORTATIONS OF WHEAT AND WHEAT FLOUR, 1862 to 1906.
Wheat flour is reckoned at its equivalent in grain.

Year.	A. Total Quantities Imported. Unit, 100,000 cwt.	B. Total Value Imported. Unit, £100,000.	C. Quantity retained per Head of the Population.	D. Average Value of Wheat and Wheat Flour in Shillings per cwt.
1862	500	286	191 lbs.	11.44
1863	309	155	118 ,,	10.03
1864	288	135	109 ,,	9.37
1865	258	124	97 ,,	9.61
1866	294	168	110 ,,	11.43
1867	391	285	144 ,,	14.58
1868	365	249	134 ,,	13.64
1869	444	233	166 ,,	10.50
1870	369	196	132 ,,	10.62
1871	444	268	158 ,,	.12.07
1872	476	303	168 ,,	12.73
1873	516	344	180 ,,	13.33
1874	493	309	170 ,,	12.53
1875	595	324	203 ,,	10.89
1876	519	279	176 ,,	10.75
1877	635	407	212 ,,	12.82
1878	597	342	197 ,,	11.46
1879	730	400	239 ,,	10.95
1880	685	393	222 ,,	11.47
1881	713	407	229 ,,	11.42
1882	808	449	257 ,,	11.11
1883	851	438	269 ,,	10.30
1884	669	301	210 ,,	9.00
1885	823	337	256 ,,	8.19
1886	670	261	207 ,,	7.79
1887	802	314	245 ,,	7.82
1888	804	315	244 ,,	7.82
1889	789	311	238 ,,	7.88
1890	824	327	246 ,,	7.94
1891	895	396	265 ,,	8.85
1892	956	371	281 ,,	7.76
1893	938	308	273 ,,	6.57
1894	967	268	277 ,,	5.54
1895	1,073	302	305 ,,	5.63
1896	996	309	279 ,,	6.21
1897	887	330	247 ,,	7.44
1898	944	377	259 ,,	7.99
1899	985	330	267 ,,	6.71
1900	986	334	266 ,,	6.78
1901	1,011	334	270 ,,	6.60
1902	1,079	360	288 ,,	6.67
1903	1,167	397	309 ,,	6.80
1904	1,182	415	310 ,,	7.02
1905	1,142	413	296 ,,	7.23

The following table contains numbers for continuing the diagram. •

Year.	A.	B.	C.	D.
1906	1,127	395	290 lbs.	7.01
1907	1,156	440	295 ,,	7.61
1908	1,091	454	275 ,,	8.32
1909	1,132	516	284 ,,	9.12
1910	1,191	497	296 ,,	8.35
1911	1,120	442	276 ,,	7.89
1912	1,237	520	301 ,,	8.41
1913	1,225	502	296 ,,	8.20

The general characteristics of a series in time are to be found in its trend and in the nature of its fluctuations, and such series may be classified as follows :—

(a) With trend, in constant or gradually changing direction, Trend and and no fluctuations. Statistics of the population fluctuations. of a country are generally in this class.*

(b) With *random* fluctuations ; that is, fluctuations of such a nature that when a movement (up or down) is recorded in a year it does not lead to any forecast as to whether the movement in the following year will be up or down. Ex. Annual statistics of rainfall.

(c) With *compensating* fluctuations; that is, when an upward movement in one year is generally compensated by a downward movement in the following. Birth, death and marriage rates frequently show such compensation.

(d) *Undulatory;* that is, when after a maximum or crisis downward movements follow one another for some years till a minimum is reached and then there are successive upward movements. General price statistics, and indeed that great mass of records which is related to the so-called commercial cycles, are of this nature.

(e) *Periodic ;* that is, when every ten years or twelve months, or some other period, the sequence of ups and downs is repeated in the same order and (in some cases) the magnitude of the fluctuations is repeated. A seasonal example is given on pp. 159 *seq.* below.

In (b), (c), (d), and (e) a trend may be combined with the

* There are also series where the records are equal over several years and then move abruptly to another level and there remain for a time. Standard time-rates afford an example of this kind.

fluctuations. We may also have random or compensated fluctuations superimposed on an undulatory movement and a trend; ripples on the waves of a rising tide. When we have a time series of records, it is very important to consider the general nature of the trend and fluctuations shown, in order to form a judgment of the near future. If fluctuations are seen to be random and violent, we shall not be disturbed by a low record and believe some remedial measures to be necessary. In the case of compensating fluctuations, we shall anticipate a high value after a low one. If the series is undulatory we shall be prepared for a deferred recovery after the figures have once broken from a high value.

3. Comparisons of Series of Figures.

A. Before proceeding to the study of the next diagram, it will be well to define more exactly what is our object in comparative studies of figures, and to consider the means at our disposal.

When dealing with two series of similar quantities such as the course of trade or population in two countries, we wish to see the general rate of progress (as can be done by *Quæsita in* smoothing the curve), the years of special increase, *comparisons.* the dates of maximum and minimum, in fact to compare the three things that the eye can see—the increase, the rate of increase, and the dates of change of rate of increase. The most obvious way to do this is, to take the same scale and base line for both countries and the same unit of measurement; but this method does not take us all the way. We can judge differences, it is true, and the additions in all the years in both countries, and we can see the highest and lowest points and *dates* of change of rate of increase; but we cannot compare rates of increase. It is not easy to judge ratio, though a rough guess at it is possible. Thus if the trade is very different in magnitude in the two countries, equal absolute increments will mean very different relative increments, and it is difficult to be always on one's guard.

The remedy for this is to alter the arrangement of scales. Make a second figure, in which the unit shall be not a sum of money, but a percentage : let 1 per cent. of Eng- *Percentage* land's trade, say in 1850, be the unit for the *scales.* English line; and 1 per cent. of the trade of Germany, at the same date, for the German line. In other words, express the trade of both countries as percentages of their value in a given

year, and draw lines to represent these percentages. Alongside the diagram two or more scales can be placed showing the absolute amounts of the trade of each country. Then the rates of increase will be comparable, equal increments representing equal percentages of the trade of each country in 1850; and, in addition, the dates at which either country gained ground relatively to the other can be easily picked out. The question whether absolute rates or relative rates should be studied is a very common one in statistics. Sometimes the absolute

Absolute or relative progress. magnitude should be known, as for instance when we want to estimate the effect of measures which will affect the well-being of special classes, or the trade of special countries; sometimes the relative rate, as when we want to watch the progressive increase of different industries, or to be on our guard as to future competitors. The two studies generally require two different diagrams though they may represent the same numbers.

It will be seen that the chief difficulty lies in the choice of the year in which the quantities are to be equated; this must be decided by the nature of the argument which the diagram is to illustrate.

We may compare the following numbers—

Year	-	1880	1890	1900
A -	-	220	440	330
B -	-	160	240	400

in three ways, shown in the diagrams on p. 151.

In Figure 3 the fluctuations are seen as percentages of the values at the last date, and are thrown into better proportion than in Figure 1. It is frequently the case that the equating of quantities at the most recent date throws what are often small beginnings into their right proportion when viewed from the modern standpoint. The statements that the values in 1880 were 40 and 67 per cent. respectively of the corresponding present values, is in better perspective than the statement that the values in 1900 were 250 per cent. and 150 per cent. of the corresponding values in 1880; but circumstances must decide in each case which method is to be adopted.

1. Expressed as percentages of values in 1880.
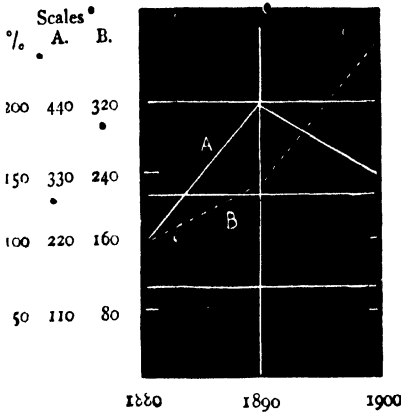
Scales
%   A.   B.

200   440   320

150   330   240

100   220   160

50   110   80

1880   1890   1900

2. Expressed as percentages of values in 1890.

Scales
%   A.   B.

150   660   360

100   440   240

50   220   120

1880   1890   1900

3. Expressed as percentages of values in 1900.

Scales
%   A.   B.

150   495   600

100   330   400

50   165   200
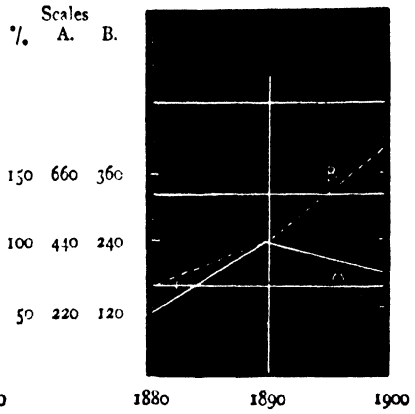
1880   1890   1900
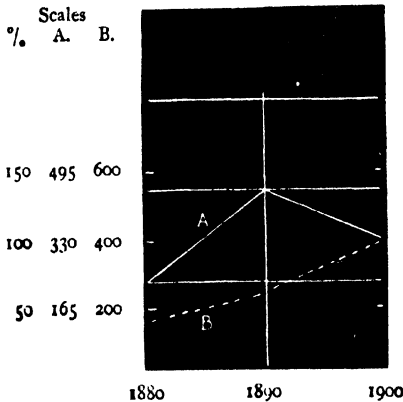
. These points are fully illustrated by the annexed diagrams, the object of which is to analyse the progress of our trade with our colonies and with foreign countries, especially Germany. The first figure shows the total im- ports and exports, and the parts of each which are colonial and foreign, the scale in millions of pounds being

Illustration from trade with Germany.

the same for all the lines. A line is also given for imports from Germany, Holland, and Belgium; these are grouped together, because it was not possible till 1904 to distinguish in the returns

<div align="center">

IMPORTS AND EXPORTS, 1862-1905.

Unit in all columns, £100,000.

</div>

	Total Imports.	Total Exports including Re-exports.	Exports to British Possessions.	Exports to Foreign Countries.	Imports from British Possessions.	Imports from Foreign Countries.	Imports from Germany, Holland and Belgium.
1862	2,257	1,662	454	1,207	653	1,604	279
1863	2,489	1,969	550	1,419	847	1,642	283
1864	2,749	2,126	557	1,569	937	1,812	332
1865	2,711	2,188	515	1,673	728	1,982	364
1866	2,953	2,389	572	1,817	722	2,231	388
1867	2,752	2,258	534	1,724	607	2,144	373
1868	2,947	2,278	537	1,741	670	2,277	379
1869	2,955	2,370	519	1,851	704	2,250	405
1870	3,033	2,441	554	1,887	648	2,384	409
1871	3,310	2,836	556	2,280	729	2,531	469
1872	3,547	3,146	656	2,490	794	2,753	455
1873	3,713	3,110	711	2,399	810	2,903	463
1874	3,701	2,977	779	2,197	822	2,879	494
1875	3,739	2,816	767	2,050	844	2,895	515
1876	3,752	2,568	701	1,866	843	2,908	516
1877	3,944	2,523	758	1,766	896	3,049	590
1878	3,688	2,455	720	1,735	779	2,908	575
1879	3,630	2,488	665	1,823	789	2,840	543
1880	4,112	2,864	815	2,049	925	3,187	616
1881	3,970	2,971	867	2,104	915	3,055	582
1882	4,130	3,067	923	2,143	994	3,136	658
1883	4,269	3,054	904	2,150	987	3,282	692
1884	3,900	2,960	883	2,077	958	2,942	646
1885	3,710	2,715	885	1,860	844	2,866	638
1886	3,499	2,690	822	1,867	819	2,680	609
1887	3,622	2,813	823	1,990	838	2,784	646
1888	3,876	2,986	917	2,068	869	3,007	684
1889	4,276	3,156	908	2,248	973	3,304	715
1890	4,207	3,283	945	2,337	962	3,245	694
1891	4,354	3,091	933	2,158	995	3,360	716
1892	4,238	2,916	812	2,104	979	3,259	715
1893	4,047	2,771	786	1,986	919	3,128	720
1894	4,083	2,739	786	1,952	940	3,143	716
1895	4,167	2,858	761	2,098	957	3,210	729
1896	4,418	2,964	907	2,057	933	3,485	761
1897	4,510	2,941	871	2,071	941	3,569	760
1898	4,705	2,940	901	2,038	998	3,708	786
1899	4,850	3,295	943	2,352	1,069	3,781	834
1900	5,231	3,544	1,021	2,523	1,096	4,134	861
1901	5,220	3,479	1,132	2,347	1,057	4,163	897
1902	5,284	3,492	1,176	2,317	1,069	4,215	950
1903	5,426	3,604	1,195	2,409	1,137	4,289	973
1904	5,510	3,710	1,208	2,502	1,200	4,310	962
1905	5,650	4,076	1,227	2,849	1,279	4,372	990

from the two latter home manufactures from German goods in transit. It is not clear from this diagram which part of our imports has increased most rapidly. The three lines are, therefore, redrawn in the second diagram, on a percentage scale,

all the values being expressed as percentages of the corresponding values in 1905. It is now seen that imports from foreign countries and from our colonial possessions and India have marched together except during the period of the cotton famine, but the trade from Germany, etc., has increased more rapidly than either. If we had equated the quantities in 1862, the German line would have far outpassed the others by 1905; but the impression given would be erroneous as regards absolute quantities, for the increase was only £71,100,000 for the one, while it was £277,000,000 for all foreign countries. The remaining diagram shows the relative rates of increase for Germany, Holland and Belgium, and the British possessions respectively, since 1870.

The International Institute of Statistics has considered the possibility of standardising historical diagrams for comparison, and resolved at its meeting in 1911 that the average of the figures for the years 1901–10 should be taken as the standard and that this average should be represented by a vertical height equal to the horizontal measurement that represented thirty years. Diagrams drawn on this standardised scale can then readily be compared with one another whatever quantities they represent. It is not intended to prevent other comparisons being made (as, for example, those on the diagram facing p. 146), nor diagrams that represent series all expressed in the same units (£ or tons) being drawn with the same natural unit. The intention is that the standard should be adopted as the only form where there is no reason to the contrary, and as an alternative form in other cases. Comparison, especially of international statistics, will be greatly facilitated if these rules are followed.

B. Series of figures are often compared graphically with a view to discovering or illustrating causal relations. In such cases we do not study relative growth only as in the last diagram discussed, but look throughout the period for any signs of resemblance in rates of growth, dates of maxima and minima, or synchronism in any changes. The methods by which such comparisons are made are difficult, and need careful analysis. For instance, we may wish to consider whether an increase of the allowance for outdoor relief is connected with an increase of pauperism. In this case one line will represent money, the other the number of persons, and there

is no common unit; we need not calculate percentages, but having chosen any scale for money, we can make equality in any year by a simple adaptation of the scale for number. We shall wish to learn first, whether an increase or decrease ⸱of money occurred at, or just before, an increase or decrease in number; and secondly, whether the greater the increase of one the greater the increase of the other. In order to show direct connection, we shall try to make one line lie as nearly as possible over the other.

Draw a preliminary diagram in which both lines are entered on any scales; this will suggest the resemblances to.be tested.

Construction.   Notice in what period the fluctuations are greatest; this in general should be the period to be taken, for it is here that the causal relations have had most play. If any other period is chosen for any special reasons, these should be made clear, for otherwise a critic may legitimately object that it is only in this period that the connection is distinct. There would be little difficulty in finding short periods in any two curves where the fluctuations synchronised. Take the averages of both money and of number over the period chosen, and draw a second diagram in which the scale for number is chosen by making this average for number equal to the corresponding average for money. Any correspondence between the two lines can be at once detected.

There are many cases when the changes in the magnitudes which we regard as the causes are in the opposite direction

Inverse   to those in the magnitudes which we regard as the
relations.   effects. For instance, if we are comparing trade improvement with the number of unemployed, and make the construction just described, the maxima of the first line would synchronise with the minima of the second. Greater clearness can be obtained by inverting one of the diagrams, plotting out⸱ the number employed instead of that unemployed, and then the changes should be in the same sense in both lines.

In the above construction the lines will only lie one over the other throughout their fluctuations, if the changes in one

More complex   quantity are in strict proportion to the changes
relations.   in the other, if an increase of 10 per cent. above the average, for instance, in the allowance for outdoor relief corresponded to one of 10 per cent. in the number of paupers. It is very rare that such a simple relation is found; all we can see

in general is that the maxima and minima occur at the same dates, that the fluctuations agree throughout in sense in both series, and that the greater fluctuations in the one correspond to the greater fluctuations in the other.

Diagrams may often be used to suggest correlation between two series of figures, and this indeed is one of their chief merits, and they may be used to *illustrate* arguments on the subject, but at this point their utility ends, for they cannot be made to *prove* much. Causal relations are very difficult to establish, and the original figures must be critically consulted when theories are to be brought to the test. Use of diagrams.

We have not yet exhausted the power of diagrams for making such comparisons, but the following method must be applied only with great caution. Suppose that we wish to ascertain whether an increase of 1 bushel in the quantity of wheat to be bought for a sovereign corresponded to an increase of 1·5 in the marriage rate per 1000, or any such strict numerical proportion. Draw a diagram representing the quantities of wheat, take the average for the period chosen for comparison, and write the scale so as to read 1, 2, 3 . . . bushels *above or below the average. Draw no base line.* Now enter a line to represent the excess or defect of the marriage rate from its average in the chosen period, on a scale such that 1·5 in excess is represented by the same vertical distance as 1 bushel. The closeness of the two lines would test to what extent the theory was valid. The danger of this method is, that with no base line there is no possibility of judging the amounts of the changes relative to the totals. The insertion of the necessary two base lines would confuse rather than aid. More exact method.

It is clear from the preceding analysis that, by the choice of scales and base lines, the points at any two dates may be made to coincide on any number of accurately drawn lines representing series of figures.

The preceding paragraphs are completely illustrated by the adjoining diagram.

In Figure I are given lines representing the price of wheat in shillings per quarter, the total of values of exports and imports divided by the population, and the marriage rate per 1000. The scales chosen are simply those which are easiest to use, and throw the lines into proper relief. Illustration of method.

MARRIAGE RATE, TOTAL EXPORTS AND IMPORTS PER HEAD OF POPULATION, AND AVERAGE PRICE OF WHEAT PER QUARTER.

Year.	Marriage Rate.	Total Exports and Imports per Head.			Average Price of Wheat per Quarter.	
		£	s.	d.	s.	d.
1860	17.1	13	0	8	53	3
1861	16.3	13	0	3	55	4
1862	16.1	13	8	0	55	5
1863	16.8	15	2	7	44	9
1864	17.2	16	8	7	40	2
1865	17.5	16	7	5	41	10
1866	17.5	17	14	5	49	11
1867	16.5	16	9	6	64	5
1868	16.1	17	0	6	63	9
1869	15.9	17	3	9	48	2
1870	16.1	17	10	3	46	10
1871	16.7	19	9	6	56	8
1872	17.4	21	0	0	57	0
1873	17.6	21	4	2	58	8
1874	17.0	20	11	0	55	8
1875	16.7	19	19	4	45	2
1876	16.5	19	0	10	46	2
1877	15.7	19	5	5	56	9
1878	15.2	18	2	1	46	5
1879	14.4	17	16	10	43	10
1880	14.9	20	3	3	44	4
1881	15.1	19	17	5	45	4
1882	15.5	20	8	10	45	1
1883	15.5	20	13	2	41	7
1884	15.1	19	4	1	35	8
1885	14.5	17	16	9	32	10
1886	14.2	17	0	10	31	0
1887	14.4	18	11	7	32	6
1888	14.4	18	12	1	31	10
1889	15.0	19	19	9	29	9
1890	15.5	19	19	7	31	11
1891	15.6	19	14	0	37	0
1892	15.4	18	15	6	30	3
1893	14.7	17	14	9	26	4
1894	15.1	17	11	9	22	10
1895	15.0	17	19	3	23	1
1896	15.8	18	14	1	26	2

The points in each scale for the same years are over one another, but the scales differ. The base lines need not coincide.

We can see at a glance whether there is resemblance between the courses of these figures. There is at any rate a general **Marriage rate and trade.** correspondence between the fluctuations of trade and of the marriage rate since 1870, and possibly earlier. There are points of likeness between wheat prices and trade; in 1870–73 both rise together, and fall in 1873–75; both rise in 1876–77, fall in the following two years, and then

rise again; both fall from 1881 to 1886 and then rise. There are also many cases in which the motions do not agree, especially 1862–64, and 1887–89.

' If we look now at the price of wheat and the marriage rate, which in the earlier part of the century used to be closely related, the one rising when the other fell, we see *Marriage rate* that there is no great resemblance either in this *and wheat.* or 'the contrary sense. In 1860–62 and in 1862–64 wheat rose and fell, while the marriage rate fell and rose; wheat rose in 1865–67, while the marriage rate was first stationary and then fell a little; then it continued to fall in 1868–70, though wheat was falling also; in 1870–80 the marriage rate shows one long, wheat two short, fluctuations. Since 1880, in years in which wheat fell, the marriage rate in general fell also and *vice versa*.

Let us consider for a moment the possible links of connection between these phenomena. When wheat was the chief object of expenditure of the working class, its *Connecting* price was the chief thing for them to consider; *links.* and so when wheat rose the marriage rate fell. On the other hand, now that wheat is cheap and wages higher, a change in the price of the loaf is only of great importance to a minority; it is now the general prosperity of the country, well indicated by the condition of foreign trade, that raises the marriage rate.

When exports and imports are increasing in value, trade is stimulated, and in spite of rising prices, marriageable people are sanguine that the prosperity will remain and the prices fall; but when the prices fall, so do the profits and incomes, and marriageable people are more prudent. For these reasons we may expect the marriage rate and foreign trade lines to resemble each other.

Now the increase of the marriage rate corresponding to an 'inflation of trade, and an inflation of trade to a time of rising prices in general, we shall find the price of wheat in particular, which is connected with the course of prices in general, rising when trade is inflated and falling when it is depressed, and therefore rising and falling with the marriage rate. But since the price of wheat is influenced also by special causes, it will not always correspond to the state of trade, and still less to the marriage rate, with its former tendency to opposite variations.

There is no need then for surprise that the curves marriage rate and trade correspond; that wheat and trade correspond,

but less closely; and that wheat and marriage show a double
tendency. The correspondence between marriage and trade is
investigated on the diagram. That between wheat and trade
should be done on an identical method. Marriage and wheat
should be compared twice on different plans : first for direct
correspondence, and then by redrawing the wheat curve with
its base line at the top for inverse correspondence.

To effect the comparison between the course of trade and
the marriage rate, the following steps are taken. On examining
Construction of the two curves on the first figure, it is seen that
diagram.   the resemblance does not begin before 1869;
the parts of the curves since 1869 should therefore be brought
into close correspondence. The average marriage rate, 1869–96,
is 15·5, and average imports and exports per head, £19. The
marriage curve is drawn in the ordinary way; then with the
help of a sliding scale the trade curve is put in, so that with
the same base line £19 falls on the 15·5 line in Figure II.

The result is that the curves are seen to rise and fall at the
same dates, but not to the same extent ; for, while the lines
keep nearly parallel from 1873 to 1879, the falls from the
maximum being equal, after 1879 the trade line fluctuates
further above and below its average than the marriage rate does.

It remains to test graphically whether the changes are
proportional to one another. An equation of scales may be
Final   obtained by equating the mean deviation (£1·04)
Comparison.   of imports and exports from their average 1869–96,
with the mean deviation of the marriage rate (·72) from its
average in the same period ; or roughly taking the same vertical
scale to represent £1 of imports and ·7 in the marriage rate.
This is making the hypothesis that a change of £1 in the total
trade per head synchronises with a change of ·7 in the marriage
rate per thousand. The scales so chosen are marked above
and below the common average line in Figure III.

It is now seen that the fluctuations since 1870 lie more
closely together in the two curves, but that this closeness has
been obtained by the partial sacrifice of the years before 1870.
A yet shorter period, 1879–93, would show a very close agree-
ment ; but so special a selection would vitiate any general
argument.

Our conclusion is, that since 1870 the causes which affect
foreign trade have also affected the marriage rate at the same

dates and in the same sense, and that the more marked the effects on the one, the more marked are the effects on the other also, but that there is no law of simple proportion between them.

'Instead of making comparison of the deviations from the average of a period, it is legitimate and often advantageous to measure the deviations from a smooth curve, whether obtained from moving averages or by some other method. We are then ignoring the causes which have a gradual and permanent effect, and comparing the short-period fluctuations. We return to this subject below (Part II, end of Chap. VI).

Note.—The relations tested in Figure II may be represented by the equation $\frac{x}{a} = \frac{y}{b}$, and in Figure III by $\frac{x-a}{y-b} = c$ (a constant), where x and y stand for the value of trade and the marriage rate, and a and b for their average values, and c is chosen so as to make the average fluctuations of the two sets of quantities equal. By the method of least squares c could be chosen so that the correspondence should be closer than with the value given by the calculation in the text.

4. PERIODIC FIGURES.

We now come to the consideration of periodic figures; that is, of figures which within a given period, in a year for instance when returns are monthly, reach maxima and minima at assigned times, and show fluctuations recurring with regularity in successive periods. In physical phenomena, such as the sunrise, the same daily numbers will represent the phenomena, almost without change, year after year. In the case of the tides we find a link between the more rigid annual curves of seasonal phenomena, and the less marked periods of social statistics; for the tides are subject to separate influences with periods of 24 hours, 24 hours 50 min., 29 days, 1 year, and others, and the effects of these influences are often masked one by the other. In the weekly figures of the Bank of England, Jevons discovered monthly, quarterly, and annual periods.*

In social and industrial statistics we usually find an annual period, combined with a general slow movement upwards or

Periodic figures.

(* See *Investigations in Currency and Finance.*

downwards, and confused by an irregular period of about ten years, due to alternate inflation and depression of trade. The influences of these three movements on the resulting numbers can be investigated, and the general methods of examining periodic figures fully explained by the complete discussion of one example, viz., the monthly returns of want of employment of the Friendly Society of Ironfounders. For another example the reader is referred to Jevons' essay, *On the Frequent Autumnal Pressure in the Money Market;* * and for an exercise, to the monthly gazette wheat prices, where the gradual change of the shape of the annual diagram can be traced in relation with the increasing influence of harvests in all the quarters of the globe.

These figures are specially suitable for showing graphically a double period, and the influences of rapid annual fluctuations General features and general movements of longer period on each of the figures. other. Looking at the table on p. 161 along the lines for the several years, we shall see that there is always a fall in the middle of the year. Looking down a vertical column under any month, it will be seen that there is no generally marked tendency towards increase or diminution, for high and low numbers occur in the first as well as the last few years. The most noticeable feature of these figures is the alternation of groups of years of high and of low numbers. Percentages above 10 will be found in 1857-58, 1861-63, 1866-70, 1876-81, 1884-87, and 1892-93. Let us choose for examination the period 1866-70. The figure for January 1866 is below the Januaries of previous years; those of February, March, and April are also low; from May to September the figures are greater than those of 1865 or 1864; from October to December they are greater than those of 1863, 1864, or 1865; in December 1867 they are greater than any previous year. Most of the figures for 1868 are higher than in the nine previous years; but from September 1868 onwards the figure is lower than the one twelve months earlier till September 1872. This wave of unemployment then lasted from May 1866 to September 1872.

Now let us watch the seasonal influence. In 1866 there was no fall in the summer except in April, and there was a very

---

* See *Investigations in Currency and Finance.*

PERIODIC FIGURES.

NUMBER OF UNEMPLOYED IRONFOUNDERS, expressed as percentages of estimated total number of members, month by month : calculated from figures given in the Annual Report of the Friendly Society of Ironfounders, 1894.

Year.	Jan.	Feb.	Mar.	April.	May.	June.	July.	Aug.	Sept.	Oct.	Nov.	Dec.	Average for Year.
1855	11.1	14.1	14.0	12.5	10.0	9.9	8.7	8.7	6.8	7.7	8.8	12.0	10.4
1856	10.9	12.6	12.2	10.0	9.4	7.5	6.9	7.3	6.9	8.1	8.7	9.9	9.2
1857	10.1	9.5	8.7	8.7	8.1	7.3	6.8	6.9	6.2	8.0	14.0	17.7	9.3
1858	20.2	20.6	20.9	19.8	20.3	17.8	15.9	14.3	13.1	11.9	11.5	11.2	16.5
1859	10.6	8.8	6.5	5.2	4.0	4.4	3.2	3.6	3.4	3.8	4.6	5.1	5.3
1860	4.0	3.2	2.6	2.2	1.6	1.7	2.3	2.6	2.6	2.9	3.7	5.6	2.9
1861	6.0	6.9	6.5	7.9	7.8	8.4	6.9	7.9	9.5	10.7	12.4	13.8	8.7
1862	14.5	14.0	14.0	14.6	14.4	13.7	13.3	12.9	12.2	13.5	14.9	16.0	14.0
1863	15.5	13.9	13.6	11.6	10.4	9.3	8.1	7.8	7.4	6.6	5.3	5.0	9.5
1864	6.0	7.1	6.6	5.1	4.4	3.3	2.8	2.8	2.6	3.3	4.2	8.1	4.7
1865	5.4	5.3	5.3	4.6	3.4	2.9	2.6	3.1	2.7	2.6	2.3	4.9	3.8
1866	4.2	5.4	5.1	3.6	5.1	6.5	5.9	6.5	6.9	7.4	9.3	13.8	6.7
1867	12.4	13.2	15.4	16.7	14.9	14.6	14.2	13.9	15.7	16.3	18.9	22.6	15.7
1868	22.1	20.9	19.8	18.6	16.7	15.8	14.9	14.7	14.2	14.1	15.6	17.4	17.1
1869	17.3	17.1	16.8	15.6	15.2	13.6	13.3	11.8	13.1	13.6	14.8	15.3	14.8
1870	14.5	10.9	8.7	7.2	5.0	4.5	3.7	4.5	4.9	5.0	5.6	8.3	6.9
1871	7.2	5.6	3.6	2.8	1.6	1.5	1.6	1.2	.9	1.4	1.1	2.2	2.6
1872	1.1	1.1	.9	.8	1.2	.7	.9	1.0	1.3	1.8	2.6	4.1	1.5
1873	3.3	2.8	2.7	2.5	2.1	2.0	3.0	4.9	4.3	3.3	3.3	5.1	3.3
Average 1855-73	10.3	10.2	9.7	8.9	8.2	7.7	7.1	7.2	7.1	7.5	8.5	10.4	8.6
1874	4.9	3.9	3.9	3.5	4.9	3.9	3.8	3.4	3.5	3.7	3.9	5.0	4.0
1875	4.6	3.4	3.5	2.8	2.8	2.8	3.3	3.4	3.6	4.1	4.1	5.0	3.6
1876	4.9	4.9	4.9	5.4	4.8	5.2	5.7	5.8	6.4	6.4	6.2	10.3	5.9
1877	...	7.4	7.0	6.9	8.4	7.6	7.4	7.8	9.6	10.9	12.3	16.3	9.1
1878	14.0	14.3	13.5	15.3	13.3	14.6	13.6	13.2	13.3	14.0	15.7	21.0	14.7
1879	23.2	23.8	24.7	25.5	22.3	23.4	21.5	22.6	22.5	21.1	18.0	16.6	22.1
1880	15.2	12.9	11.1	10.0	10.0	9.7	9.8	10.0	10.0	9.2	9.2	10.2	10.6
1881	11.5	10.8	10.1	10.1	7.6	7.5	6.5	5.8	5.6	5.4	5.0	6.6	7.7
1882	5.5	5.2	5.3	4.5	3.6	3.8	3.2	3.4	3.6	4.1	4.4	6.0	4.4
1883	3.6	4.8	5.2	4.3	4.2	3.6	3.9	4.3	4.3	4.2	4.0	6.6	4.4
1884	6.1	6.2	5.9	6.5	6.5	6.9	6.5	7.6	8.1	7.8	9.8	10.9	7.4
1885	10.2	11.1	10.0	10.1	9.8	9.1	9.8	10.7	11.8	11.6	12.7	13.6	10.9
1886	14.1	15.0	15.2	15.5	13.4	13.1	12.1	12.7	13.6	13.9	12.7	12.9	13.7
1887	12.4	11.6	10.2	9.1	9.2	10.6	9.2	8.8	9.6	9.4	9.4	9.1	9.9
1888	7.8	7.5	6.4	6.4	5.9	5.2	5.7	5.0	5.1	4.8	3.2	3.5	5.5
1889	3.1	3.3	2.4	2.1	1.7	1.6	1.7	1.7	1.6	1.5	1.2	1.4	1.9
1890	1.3	1.3	3.2	3.1	2.8	2.4	2.4	2.7	2.7	2.7	2.7	2.7	2.5
1891	3.9	3.5	4.2	4.2	4.6	4.0	4.5	4.8	5.4	5.6	5.7	6.3	4.7
1892	7.0	7.2	7.9	8.1	7.9	7.9	7.7	7.6	9.3	11.4	10.9	12.0	8.7
1893	11.5	11.2	10.1	7.7	9.6	8.3	8.3	9.2	11.7	11.9	11.5	11.5	10.2
Average 1874-93	8.6	8.5	8.2	8.1	7.7	7.6	7.3	7.5	8.1	8.2	8.1	9.4	8.1
Average 1855-93	9.4	9.3	8.9	8.5	7.9	7.6	7.2	7.4	7.6	7.9	8.3	9.9	8.3

M *

rapid rise in December. In 1867 a fall from April to August was followed by a rapid rise for four months. There **Seasonal** is a fall from December 1867 to September 1868, **influence.** but a rise follows in October, November, and December; since the rise does not generally begin till after August, it will be seen that the general fall did not much delay the seasonal effect. In the next year, 1869, there is a fall to a lower minimum in August, but now the rise in December is very slight, next year the fall is very quick to August, but the seasonal rise is not delayed. From this it is clear that the seasons had their effect throughout the fluctuation except in the opening year 1866, when there was no fall, and that the rises in the autumn were very much accentuated. Almost identical remarks would apply to the period August 1875 to May 1881. In what month was the condition of employment 1867–70 at its worst? The greatest figure given is 22.6 per cent. in December 1867, but unemployment in December is generally greater than in any other month, and the figures for any of the following six months may be more unusual; the determination of the exact date will be best shown by diagrams. It may be mentioned that most of these remarks were suggested by Mr. Hey, the former secretary of the Iron-founders' Society, who drew up these figures.

If we now turn to the diagram, the following facts may be. noticed. The thick line showing the annual average percent-**The story from** ages shows a downward tendency till 1857, fol-**the diagram.** lowed by an abrupt rise and fall in 1858–6₃ then two years' rise to its original height, returning to a minimum in 1865; the next wave covers seven years, and is marked by an extraordinarily sharp rise in 1867, and a very low minimum in 1872. The exceptional condition of trade in 1872 could not last, but the rise is very gradual to 1876, when the next cycle of trade is marked again by a six years' wave; the rise is not so steep as in the former fluctuation, but lasts longer, and a higher point is reached : the fall is at about the same angle, and the minimum in 1882 is about the same as that in 1865. The next wave came before it appeared to be due, and lasted seven instead of six years, but was much more moderate, and again the rise was sharper than the fall. The minimum of 1889 did not endure, and the figure ends with a suggestion that the maximum will be in 1894, but only at a moderate height, and the next minimum might be expected in 1898

or 1899, if causes similar to those which influenced earlier trade depressions were still acting. It may be found, in fact, from the Board of Trade returns, that, taking all the trade unions who made returns together, the maximum month was December 1892, and the maximum year was 1893; after this the fall is regular to 1897, and a trifling rise in 1898 is followed by a very low figure for 1899.*

In Figure 5 the diagram is inverted and greatly compressed, showing now the percentage employed. If the period 1876–82 is cut off by two vertical lines, readers may see how great were the amounts of labour lost to the country and wages to the members of the Ironfounders' Society in those years. These figures show a want of employment due to special causes in this Society more than twice as great as in other Unions whose returns are available for the same period.

In Figure 5 the annual averages are smoothed by the method explained above (pp. 136–7), a seven-yearly average † being taken to correspond to the general wave length. It will be seen that there is no very marked tendency up or down in the thirty-nine years, and that the smooth line is never far from the general average of employment, 91·7.

The comparison of this diagram with that illustrating exports (p. 134) is very instructive. Some of the results may be thus exhibited :—

DATES OF		DATES OF	
Minima of Exports.	Maxima of Unemployment.	Maxima of Exports.	Minima of Unemployment.
1862	1858 and 1862	1866	1865
1868	1868	1872	1872
1879	1879	1882	1882 or 1883
1886	1886	1890	1889
1894	1893		

The figures may also be compared graphically by the methods of the previous or following sections.

The averages for the nineteen Januaries, nineteen Februaries, etc., in the years 1855–73, and similar averages for the years 1874–93, and the whole period are given in the table and exhibited in Figures 2, 3, 4. *Measurement of seasonal influence.*

---

* See *Annual Abstract of Labour Statistics*, 1895, p. 73, for various methods of treating these figures similar to those here discussed.

† For smoothing and studying periodic curves, see Professor Poynting's paper in *Statistical Journal*, 1884, and Professor Moore's in 1915.

When we calculated the annual averages just discussed we eliminated by that process the seasonal fluctuations; by this new series of averages we eliminate the influences of particular years. If we took, for instance, all the November numbers out of a series of figures totally uninfluenced by the seasons, if such could be found, and compared these with the general average for all months, we should in the long run find just as many instances above as below this average; but if the figures were influenced by the seasons, we should find a considerably greater number above than below, or *vice versa*. The greater the . seasonal influence, the greater would be this excess or defect. Averaging numbers in this way tends to eliminate the non-seasonal causes, for by hypothesis the excesses and defects due to them will in the long run balance one another; and except by averaging these cannot be eliminated, unless they can be actually calculated. The excess of the November average above the general average will be greater than that of October, if the seasonal causes exert more influence towards excess in the former than in the latter month, and the curve which shows these averages will show a resemblance to that which would be obtained if the non-seasonal causes were absent. It will be only a resemblance for two reasons : first, because in the comparatively short series of years with which we are generally obliged to be content, a very effective non-seasonal cause will leave its mark on the average, as may be seen in the table on p. 161; secondly, because seasonal and non-seasonal causes are often not independent; a depression of trade is accentuated by a sharp winter; a bad season in a year of bad trade may increase the want of employment greatly and suddenly, while a good summer in a prosperous year may reduce it almost to zero. In the case we are considering the interaction of causes tends to exaggerate the seasonal maximum and diminish the minimum; in other cases a compensating effect might be found.

In Figures 2, 3, 4 the curve for the latter half of the year is prefixed to that of the calendar year, because the character of the yearly waves is seen most clearly from minimum to minimum. It may be noticed that the wave in Figure 3 is less definite in shape and has a smaller rise and fall than that of the earlier period shown in Figure 2; it would appear that the seasons are losing their influence.

If there is a definite annual period, that represented by

Figure 4, it may be expected that a figure of a shape similar to this—

will be repeated annually in Figure 1; it is shown well in 1864, 1882, and other years.   In the great majority of cases the yearly maximum is reached in December or January; at   The annual the end of 1858 the maximum is absent, but is   wave. replaced by a break in the rapidity of the fall; at the end of 1860 there is a rise, but the spring fall following is checked by the general upward trend; similar remarks apply to all the great fluctuations.   There is no doubt that right along the line we find at nearly equal intervals these pointed crests above the line of averages.

The minima are not so conspicuous, for the pointed shape is absent, trifling causes bring them near the smoothed line, and they are easily masked by a general fall or are absent because of a general rise.   In 1861, however, there is a distinct minimum in spite of the strong upward tendency; the minima are very conspicuous throughout the fluctuation of 1865–70; and from 1859 to 1888 the minima are fairly marked, except in 1876, 1880, and 1881.

The following figures show the effect of a stationary, rising, and falling average annual rate on the shape of the seasonal wave :—

a. Seasonal wave on stationary line of averages.

b. Seasonal wave superimposed on rising line of averages.
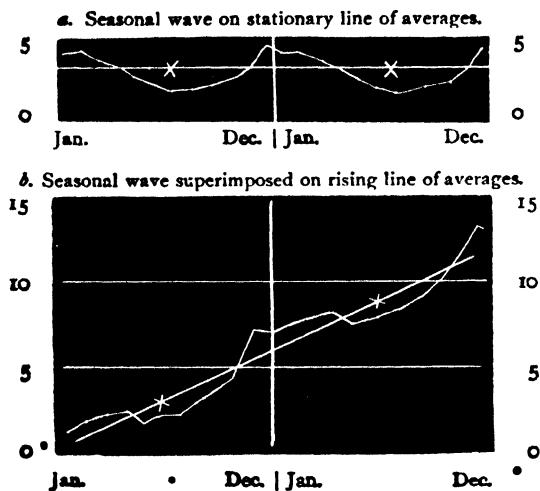
c. Seasonal wave superimposed on falling line of averages.
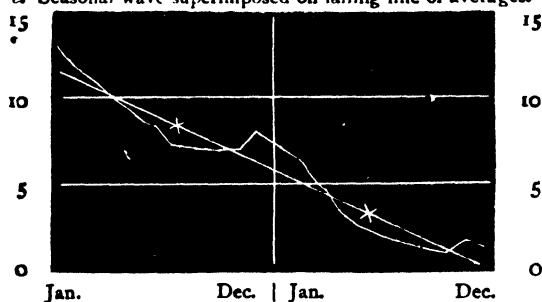
These figures are drawn by adding or subtracting the average monthly differences from the general average

$$\left(\text{viz.} \quad \begin{array}{cccccccccccc} \text{Jan.} & \text{Feb.} & \text{Mar.} & \text{Apr.} & \text{May.} & \text{June.} & \text{July.} & \text{Aug.} & \text{Sept.} & \text{Oct.} & \text{Nov.} & \text{Dec.} \\ +1\cdot1 & +1\cdot0 & +\cdot6 & +\cdot2 & -\cdot4 & -\cdot7 & -1\cdot1 & -\cdot9 & -\cdot7 & -\cdot4 & 0 & +1\cdot6 \end{array}\right)$$

month by month to or from the positions shown on the straight lines joining the annual averages. On a rising line the spring fall tends to become horizontal and the autumn rise steeper; on a falling line the spring fall becomes more rapid and the autumn rise is checked.

If this seasonal wave, added to the slower long-period changes, were the complete explanation of these numbers, Figure 1 (p. 162) would be entirely composed of modifications of Figures *a*, *b*, and *c*. Figure *a* is exemplified especially in 1855–57, 1864–65, 1871–73; Figure *b* in 1860–61, 1866–67, 1877–78, 1883–85; Figure *c* in 1859, 1863, 1880–82, 1886–89.

As explained above, the two sets of causes are not independent, and these figures are not reproduced exactly; but the Elimination of resemblance is sufficiently close to make the fluctuations. following method of eliminating seasonal fluctuations partially applicable. Combine the monthly excesses and defects just given with the original numbers, by subtracting the excesses and adding the defects; this process should tend to produce a straight line thus :—

5 ———— from figure 1
.... corrected figures.
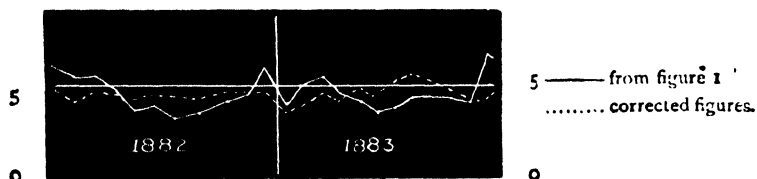
But the result is not more than a tendency, because of the unusual fall in January 1883, and it is difficult to find a perfect

example. This method is applied in Figures 6, 7, and 8 in an attempt to disentangle the seasonal fluctuations from the effects of the commercial crisis of 1872, the depression of 1879, and the turn of the tide in 1889. In Figure 6 it is seen that January 1872 was the best month relatively, though the absolute minimum was not reached till June of that year; from this it appears that January 1872 was the turning point of the great inflation, a date somewhat earlier than that generally given. The date of the maximum of 1879 is left unchanged by this process, and that of the 1889 minimum is only shifted one month.

We have still to discuss the criteria of the existence of a period. In Figure 1 the optical evidence is sufficient to suggest the annual period, but it may be doubted whether an annual fluctuation would be suggested by a diagram representing wheat prices. It is clear that if the monthly entries of any returns whatever were averaged in months over any period of years, that the averages for January, February, etc., would not be exactly equal, even if there were no seasonal influence. The following diagrams show various averages :— *Criteria of existence of period.*

Unemployed ironfounders
as before.

Wheat prices, shillings
per quarter, 1862-76.

Wheat prices, shillings per
quarter, 1877-91.

Average date of first Sunday
in month, 1881-1900.
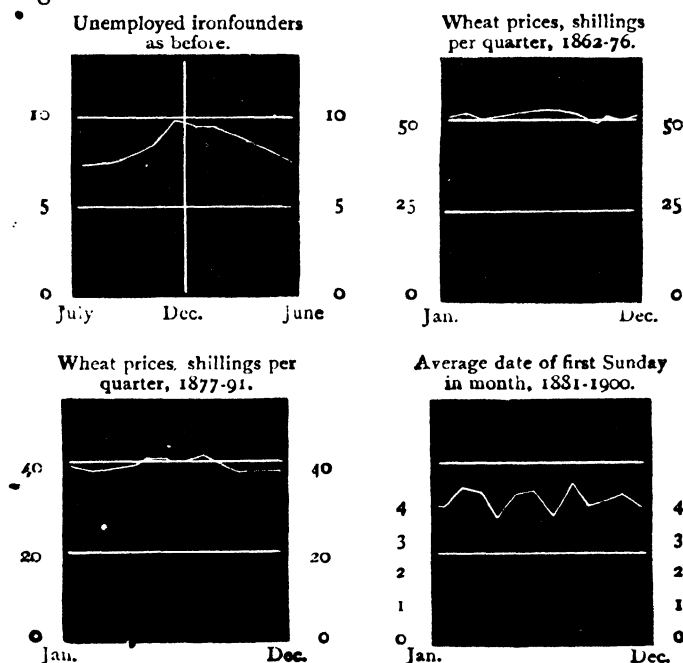
Of these the first three may be expected to be seasonal, while the last, which shows the averages of the dates on which fell the first Sunday in 20 Januaries, 20 Februaries, etc., in a series of years, certainly is not.

The following simple tests may be applied to decide this point. If the period is in any way connected with the seasons, it will correspond to some extent to the ordinary weather charts of temperature, etc., which have a single annual maximum and corresponding minimum. Phenomena affected by the weather may also be expected to show a single maximum, nearly coinciding with the maximum or minimum temperature; thus the maximum unemployed coincides with the minimum length of daylight and precedes the minimum temperature. In some cases a second subsidiary maximum may be shown, since, for example, an excessive death rate may be due to excessive cold or heat; but even in this example further analysis would probably show that the one maximum was for the old, the other for the young. Wheat prices may also show two minima due to the harvests in the two hemispheres. The " Sunday " curve just given shows four maxima, and is not seasonal. More than one maximum is evidence against periodicity till a reason is found for their existence.

The second test is to look at the serial diagram and notice how often the maximum occurs in the same month; non-periodic **Probability** causes will hide the maximum occasionally, but in **test.** the long run one month will be predominant. In Figure I the maximum occurs in March and April twice each, in February three times, in January eleven times, and in December twenty-one times. The maximum is then generally in midwinter. The minimum is not in this case so well defined.

The following table shows how this analysis can be extended:—

	Times out of 39.
The percentage of December is greater than that of the preceding November - - - -	33
The percentage of December is greater than that of the following January - - - -	28
The percentage of December is greater than that of the preceding July - - - - -	33
The percentage of December is greater than that of the following July - - - - -	30

The chances against so great a preponderance, if the seasons

had no influence, are respectively about 65,000 to 1, 160 to 1, 65,000 to 1, and 1200 to 1.* All the months may be separately tested in the same way. This method by no means exhausts the evidence, for we have only considered which of two months is the greater, and not how great is the excess when it exists. On this point the reader is referred to the paper by Professor Edgeworth, *On Methods of Statistics*, in the Jubilee Volume of the Royal Statistical Society, p. 206; this should, however, be postponed till the mathematical treatment which follows in Part II has been studied.

5. LOGARITHMIC CURVES.

A serious flaw in the graphic method as used in the previous sections is that, when we are dealing with a series of increasing figures, though the totals year by year may be increasing, we are compelled to represent equal increments on these totals by equal vertical distances; thus an increment of £20 on a total of £20 is represented by the same vertical distance as an increment of £20 on a total of £2000. Thus in the annexed figure representing exports, the fall from £52,000,000 to £42,000,000 in 1815–16 is barely noticeable, though it is a fall of 20 per cent., and was connected with very great distress in the manufacturing districts; while the fall from £305,000,000 in 1883 to £269,000,000 in 1886 attracts attention immediately, though it is one of 12 per cent. only. Again the increase of 34 per cent. which took place between 1848 and 1850 appears insignificant in comparison with that of 29 per cent. from 1870 to 1872. When we are attacking questions of causation it very frequently happens that we are more concerned to know the proportionate increase than the actual increase. When we are considering the gradual growth of our foreign trade, or when we are comparing the growth of trade of two countries, a diagram like that annexed is likely to give quite a wrong impression of the struggle that marked the early stages. We need then a diagram not of quantities, but of ratios, where equal vertical distances represent no longer equal absolute increments, but equal proportional increments, that is, equal rates of increase. By the use of logarithms a universal scale can be constructed which serves

Need for graphic representation of ratios.

this purpose. The non-mathematical student can easily accustom himself to the use of diagrams so constructed, by studying one where the actual amounts represented are entered, and noticing that whatever part of the scale he takes, doubling, halving, increasing by 20 per cent. and so on, are always represented by the same vertical distances respectively. The con-

Construction of a logarithmic diagram. struction of a diagram on this scale is as follows :— Write down the numbers in the series to be represented ; against them write down their logarithms ; on paper divided into equal squares mark at equal intervals on a vertical line numbers ascending in regular progression so as to include all the logarithms found ; mark off the dates on a horizontal line ; and on the scale thus prepared mark in the logarithms, instead of the original numbers. The table on p. 173 and the diagram facing p. 171 show the figures of imports and exports thus treated. On the right hand of figure 2 the position of the absolute numbers is given ; on the left the corresponding logarithms. A given vertical distance, 1 inch, represents the distance .301 on the logarithmic scale ; if we add this quantity to the logarithm of any number, we obtain the logarithm of twice that number for $\log a + \cdot301 = \log a + \log 2 = \log 2a$; for instance, if we increase the height of the position which represents £30 by 1 inch, we arrive at the position which represents £60. Again if we now add 1·59 of an inch, which represents ·477 on the logarithmic scale, that is $\log 3$, to the logarithm of $2a$, we obtain $\log 6a$, and we have—

$$\log 6a = \cdot477 + \log 2a = \cdot477 + \cdot301 + \log a, \text{ as above}$$
$$= \cdot778 + \log a = \log 6 + \log a ;$$

that is, we arrive at the same position on this scale whether we go by means of two separate ratios or by a single compounded ratio. Thus a diagram drawn on this principle satisfies the necessary conditions that equal vertical distances represent the same ratio in whatever part of the scale they are taken, and that any number of points can be entered without leading to inconsistencies. At the end of this section is given a table of the logarithms of 1 to 1000, correct to the third decimal place, which will be found sufficient for this purpose.

Thus on the diagram given we can find at once that imports were doubled in value between 1811 and 1836, again between

Examples of its use. 1839 and 1853, again between 1855 and 1866, and that their value increased 40 per cent. be-

tween 1886 and 1899. Or we may notice that the excess of the value of imports over that of exports was 40 per cent. of the latter both in 1850 and in 1880; that the value of imports in 1899 was thrice that of exports in 1860.

If the eye has been carefully educated to understand a diagram of this sort, if the fact that it is a *diagram of ratios, not of quantities*, is firmly impressed on the mind, then the diagram answers perfectly the object of the graphic method, that is, it gives a true instantaneous impression of a complex series of facts. If, on the other hand, it is found that a true impression is not received, through inability to take the right mental position, then diagrams on the natural scale should be employed only, always with the recollection that they may give false impressions of ratio.

It is to be noticed that no base line should be given in diagrams of this class, otherwise a false impression is at once obtained. Notice further that, while equal verti- Velocity and acceleration. cal differences represent equal ratios from any part of the diagram to any other, instead of equal increments as on the natural scale, equal degrees of slope represent equal ratios of increase (equal accelerations), instead of equal additions in equal times as on the natural scale (equal velocities). On the logarithmic scale a line rising with convexity to the horizontal shows that the ratio of increase is growing, as in imports from 1830–53 (if the line is smoothed), while concavity, as from 1854 to 1873, shows a slackening; but on the natural scale the line is convex almost throughout the two periods, showing that the actual increments were increasing all the time.

It would be useful, if space permitted, to offer several diagrams on both scales; for in many series of figures the differences exhibited by the two methods are very Useful application to index-numbers. instructive. One case may be signalised where the logarithmic scale is specially important, that is, when the original numbers represent ratios, not actual numbers. Thus in Mr. Sauerbeck's well-known diagram, drawn on the natural scale, representing his index-numbers of prices, all the numbers included are percentages of their values in certain defined years. Suppose that 100, 80, and 60 are the index-

percentage of unemployed with the marriage rate. In Fig. 1, the numbers are shown on natural scales; in Fig. 2 the averages over twenty-nine years are equated and the numbers are shown on a logarithmic scale. We might proceed as on p. 158, but to use an alternative method, the maxima and minima in various periods are written down as in the table on p. 175, and the averages of the fluctuations from maximum to minimum (expressed as percentages of the maximum) are calculated. It is found that a fluctuation of 8·4 per cent. in the number employed, in those trade, unions whose returns are accessible,* corresponds to one of 9·7 per cent. on the marriage rate. To investigate a possibly closer correspondence, assume that a portion of the number employed do not influence the marriage rate, and find what part must be subtracted before this 8·4 per cent. of the total forms as much as 9·7 per cent. of the remainder; the average percentage of members of the trade unions at work in the selected period was 95·1; 8·4 per cent. of this is 7·99, which forms 9·7 per cent. of 82·4. Thus 12·7, the difference between 95·1 and 82·4, may be considered as not influencing the question, and subtracted throughout before logarithms are taken. This process would be replaced on the natural scale by equating the averages of two series, and drawing one base line so far below the other that average fluctuations would be represented by the same vertical distance for both series; which process is exactly equivalent to that adopted on p. 158. Expressed algebraically, we are now investigating the equation—

Equation of fluctuations.

$$\log (y - c) - \log x = k, \text{ a constant,}$$

where c and k are constants to be so selected as to give the closest fit, and y and x are the quantities to be compared.

In the adjacent diagrams, Fig. 1 gives the figures in the natural scale; Fig. 2 gives them on the logarithmic scale, after they have been arranged so as to make average percentage fluctuations equal; while in Fig. 3 the shorter period, 1880–96, is treated in a method precisely similar to that of Fig. 2. The actual numbers and logarithms are given on the next page.

---

* The figures in columns 2 and 4 in the second table on the next page are taken from Mr. G. H. Wood's paper on *Some Statistics of Working Class Progress since* 1860, *Statistical Journal*, 1900, where a valuable logarithmic diagram will be found, illustrating many of the points of this section.

FIG. 1. Comparison in 1865-93.
ON NATURAL SCALE.
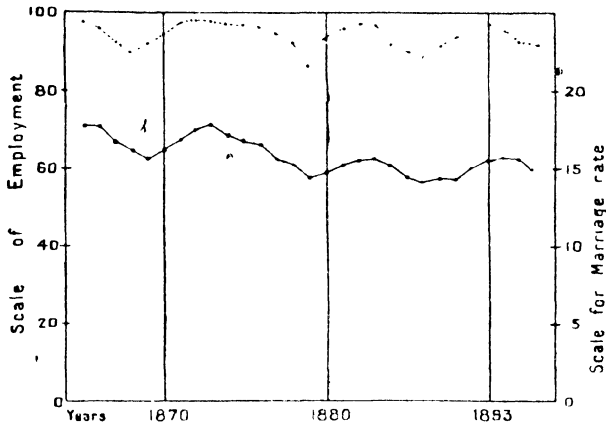
FIG. 2. The same; Logarithmic Scale.
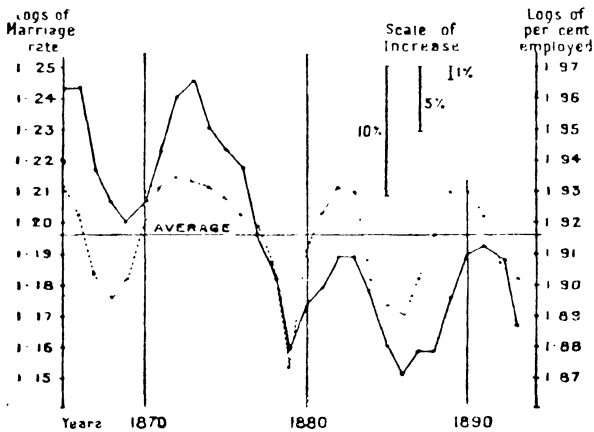
FIG. 3. Comparison in 1880-1896.

Same scale as figure 2.
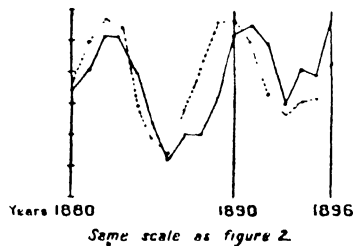
	MARRIAGE RATE PER 1,000.					PERCENTAGE EMPLOYED.			
Years.	Maxima.	Minima.	Differ-ences.	% of Max.	Years.	Maxima.	Minima.	Differ-ences.	% of Max.
1869	...	15.9	1.7	10	1868	...	91.5	7.4	7.5
1873	17.6	...	3.2	18	1872	98.9	...	11.4	11.5
1879	...	14.4	1.1	7	1879	...	87.5	10.6	10.8
1882-83	15.5	...	1.3	8	1882	98.1	...	7.6	7.8
1886	...	14.2	1.4	9	1886	...	90.5	7.4	7.6
1891	15.6	...	.9	6	1889-90	97.9	...	5.4	5.5
1893	...	14.7			1893	...	92.5		
				9.7					8.4

Average percentage employed, 1865-93, 95.1 ; 8.4 per cent. of 95.1 is 9.7 per cent. of 82.4.

Years.	Marriage Rate.	Logarithms.	Percentage Employed.	Less 12.7.	Logarithms.
1865	17.5	1.243	98.0	85.3	1.931
1866	17.5	1.243	96.9	84.1	1.925
1867	16.5	1.217	92.7	80.0	1.903
1868	16.1	1.207	91.5	78.8	1.896
1869	15 9	1.201	92.6	79.9	1.902
1870	16.1	1.207	95.7	83.0	1.919
1871	16.7	1.223	98.2	85.5	1.932
1872	17.4	1 240	98.9	86.2	1.935
1873	17.6	1.245	98.7	86.0	1.934
1874	17.0	1.230	98.2	85.5	1.932
1875	16.7	1.223	97.5	84.8	1.928
1876	16.5	1.217	96.4	83.7	1.923
1877	15.7	1.196	95.6	82.9	1.919
1878	15.2	1.182	93.7	81.0	1.908
1879	14.4	1.158	87.5	74.8	1.874
1880	14.9	1.173	94.1	81.4	1.911
1881	15.1	1.179	96.5	83.8	1.923
1882	15.5	1.190	98.1	85.4	1.931
1883	15.5	1.190	97.8	85.1	1.930
1884	15.1	1.179	92.6	79.9	1.902
1885	14.5	1.161	91.0	78.3	1.894
1886	14.2	1.152	90.45	77.7	1.890
1887	14.4	1.158	92.6	79.9	1.902
1888	14.4	1.158	95.2	82.5	1.916
1889	15.0	1.176	97.9	85.2	1.930
1890	15.5	1.190	97.9	85.2	1.930
1891	15.6	1.193	96.5	83.8	1.923
1892	15.4	1.187	93.7	81.0	1.908
1893	14.7	1.167	92.5	79.8	1.902
		Average 1.196			Average 1.916

A critical account of logarithmic curves, strongly advocating their use is given by Professor Irving Fisher, in the Quarterly Publications of the *American Statistical Association*, June 1917.

LOGARITHMS OF NUMBERS 1 TO 1,000, CORRECT TO THE NEAREST DIGIT IN THE THIRD DECIMAL PLACE.

No.	Logs.	No.	Logs.	No.	Logs.	No.	Logs.	No.	Logs.	No.	Logs.	No.	Logs.	No.	Logs.	No.	Logs.	No.	Logs.
1	0	34	1.531	67	1.826	101	2.004	134	2.127	167	2.223	201	2.303	234	2.369	267	2.427	301	2.479
2	.301	35	1.544	68	1.833	102	2.009	135	2.130	168	2.225	202	2.305	235	2.371	268	2.428	302	2.480
3	.477	36	1.556	69	1.839	103	2.013	136	2.134	169	2.228	203	2.307	236	2.373	269	2.430	303	2.481
4	.602	37	1.568	70	1.845	104	2.017	137	2.137	170	2.230	204	2.310	237	2.375	270	2.431	304	2.483
5	.699	38	1.580	71	1.851	105	2.021	138	2.140	171	2.233	205	2.312	238	2.377	271	2.433	305	2.484
6	.778	39	1.591	72	1.857	106	2.025	139	2.143	172	2.236	206	2.314	239	2.378	272	2.435	306	2.486
7	.845	40	1.602	73	1.863	107	2.029	140	2.146	173	2.238	207	2.316	240	2.380	273	2.436	307	2.487
8	.903	41	1.613	74	1.869	108	2.033	141	2.149	174	2.241	208	2.318	241	2.382	274	2.438	308	2.489
9	.954	42	1.623	75	1.875	109	2.037	142	2.152	175	2.243	209	2.320	242	2.384	275	2.439	309	2.490
10	1.	43	1.633	76	1.881	110	2.041	143	2.155	176	2.246	210	2.322	243	2.386	276	2.441	310	2.491
11	1.041	44	1.643	77	1.886	111	2.045	144	2.158	177	2.248	211	2.324	244	2.387	277	2.442	311	2.493
12	1.079	45	1.653	78	1.892	112	2.049	145	2.161	178	2.250	212	2.326	245	2.389	278	2.444	312	2.494
13	1.114	46	1.663	79	1.898	113	2.053	146	2.164	179	2.253	213	2.328	246	2.391	279	2.446	313	2.495
14	1.146	47	1.672	80	1.903	114	2.057	147	2.167	180	2.255	214	2.330	247	2.393	280	2.447	314	2.497
15	1.176	48	1.681	81	1.908	115	2.061	148	2.170	181	2.258	215	2.332	248	2.394	281	2.449	315	2.498
16	1.204	49	1.690	82	1.914	116	2.064	149	2.173	182	2.260	216	2.334	249	2.396	282	2.450	316	2.500
17	1.230	50	1.699	83	1.919	117	2.068	150	2.176	183	2.262	217	2.336	250	2.398	283	2.452	317	2.501
18	1.255	51	1.708	84	1.924	118	2.072	151	2.179	184	2.265	218	2.338	251	2.400	284	2.453	318	2.502
19	1.279	52	1.716	85	1.929	119	2.076	152	2.182	185	2.267	219	2.340	252	2.401	285	2.455	319	2.504
20	1.301	53	1.724	86	1.934	120	2.079	153	2.185	186	2.270	220	2.342	253	2.403	286	2.456	320	2.505
21	1.322	54	1.732	87	1.940	121	2.083	154	2.188	187	2.272	221	2.344	254	2.405	287	2.458	321	2.507
22	1.342	55	1.740	88	1.944	122	2.086	155	2.190	188	2.274	222	2.346	255	2.407	288	2.459	322	2.508
23	1.362	56	1.748	89	1.949	123	2.090	156	2.193	189	2.276	223	2.348	256	2.408	289	2.461	323	2.509
24	1.380	57	1.756	90	1.954	124	2.093	157	2.196	190	2.279	224	2.350	257	2.410	290	2.462	324	2.511
25	1.398	58	1.763	91	1.959	125	2.097	158	2.199	191	2.281	225	2.352	258	2.412	291	2.464	325	2.512
26	1.415	59	1.771	92	1.964	126	2.100	159	2.201	192	2.283	226	2.354	259	2.413	292	2.465	326	2.513
27	1.431	60	1.778	93	1.968	127	2.104	160	2.204	193	2.286	227	2.356	260	2.415	293	2.467	327	2.515
28	1.447	61	1.785	94	1.973	128	2.107	161	2.207	194	2.288	228	2.358	261	2.417	294	2.468	328	2.516
29	1.462	62	1.792	95	1.978	129	2.111	162	2.210	195	2.290	229	2.360	262	2.418	295	2.470	329	2.517
30	1.477	63	1.799	96	1.982	130	2.114	163	2.212	196	2.292	230	2.362	263	2.420	296	2.471	330	2.519
31	1.491	64	1.806	97	1.987	131	2.117	164	2.215	197	2.294	231	2.364	264	2.422	297	2.473	331	2.520
32	1.505	65	1.813	98	1.991	132	2.121	165	2.217	198	2.297	232	2.365	265	2.423	298	2.474	332	2.521
33	1.519	66	1.820	99	1.996	133	2.124	166	2.220	199	2.299	233	2.367	266	2.425	299	2.476	333	2.522
				100	2.000					200	2.301					300	2.477		

LOGARITHMS OF NUMBERS 1 TO 1,000, CORRECT TO THE NEAREST DIGIT IN THE THIRD DECIMAL PLACE.

No.	Logs.	No.	Logs.	No.	Logs.	No.	Logs.	No.	Logs.	No.	Logs.	No.	Logs.	No.	Logs.	No.	Logs.
335	2.525	401	2.603	467	2.669	535	2.728	601	2.779	667	2.824	735	2.866	801	2.904	900	2.954
337	2.528	403	2.605	469	2.671	537	2.730	603	2.780	669	2.825	737	2.867	804	2.905	903	2.956
339	2.530	405	2.607	471	2.673	539	2.732	605	2.782	671	2.827	739	2.869	807	2.907	906	2.957
341	2.533	407	2.610	473	2.675	541	2.733	607	2.783	673	2.828	741	2.870	810	2.908	909	2.959
343	2.535	409	2.612	475	2.677	543	2.735	609	2.785	675	2.829	743	2.871	813	2.910	912	2.960
345	2.538	411	2.614	477	2.679	545	2.736	611	2.786	677	2.831	745	2.872	816	2.912	915	2.961
347	2.540	413	2.616	479	2.680	547	2.738	613	2.787	679	2.832	747	2.873	819	2.913	918	2.963
349	2.543	415	2.618	481	2.682	549	2.740	615	2.789	681	2.833	749	2.874	822	2.915	921	2.964
351	2.545	417	2.620	483	2.684	551	2.741	617	2.790	683	2.834	751	2.876	825	2.916	924	2.966
353	2.548	419	2.622	485	2.686	553	2.743	619	2.792	685	2.836	753	2.877	828	2.918	927	2.967
355	2.550	421	2.624	487	2.688	555	2.744	621	2.793	687	2.837	755	2.878	831	2.920	930	2.968
357	2.553	423	2.626	489	2.689	557	2.746	623	2.794	689	2.838	757	2.879	834	2.921	933	2.970
359	2.555	425	2.628	491	2.691	559	2.747	625	2.796	691	2.839	759	2.880	837	2.923	936	2.971
361	2.558	427	2.630	493	2.693	561	2.749	627	2.797	693	2.841	761	2.881	840	2.924	939	2.973
363	2.560	429	2.632	495	2.695	563	2.751	629	2.799	695	2.842	763	2.883	843	2.926	942	2.974
365	2.562	431	2.634	497	2.696	565	2.752	631	2.800	697	2.843	765	2.884	846	2.927	945	2.975
367	2.565	433	2.636	499	2.698	567	2.754	633	2.801	699	2.844	767	2.885	849	2.929	948	2.977
369	2.567	435	2.638	501	2.700	569	2.755	635	2.803	701	2.846	769	2.886	852	2.930	951	2.977
371	2.569	437	2.640	503	2.702	571	2.757	637	2.804	703	2.847	771	2.887	855	2.932	954	2.978
373	2.572	439	2.642	505	2.703	573	2.758	639	2.806	705	2.848	773	2.888	858	2.933	957	2.980
375	2.574	441	2.644	507	2.705	575	2.760	641	2.807	707	2.849	775	2.889	861	2.935	960	2.981
377	2.576	443	2.646	509	2.707	577	2.761	643	2.808	709	2.851	777	2.890	864	2.937	963	2.982
379	2.579	445	2.648	511	2.708	579	2.763	645	2.810	711	2.852	779	2.892	867	2.938	966	2.984
381	2.581	447	2.650	513	2.710	581	2.764	647	2.811	713	2.853	781	2.893	870	2.940	969	2.985
383	2.583	449	2.652	515	2.712	583	2.766	649	2.812	715	2.854	783	2.894	873	2.941	972	2.986
385	2.585	451	2.654	517	2.713	585	2.767	651	2.814	717	2.856	785	2.895	876	2.942	975	2.988
387	2.588	453	2.656	519	2.715	587	2.769	653	2.815	719	2.857	787	2.896	879	2.944	978	2.989
389	2.590	455	2.658	521	2.717	589	2.770	655	2.816	721	2.858	789	2.897	882	2.945	981	2.990
391	2.592	457	2.660	523	2.719	591	2.772	657	2.818	723	2.859	791	2.898	885	2.947	984	2.992
393	2.594	459	2.662	525	2.720	593	2.773	659	2.819	725	2.860	793	2.899	888	2.948	987	2.993
395	2.597	461	2.664	527	2.722	595	2.775	661	2.820	727	2.862	795	2.900	891	2.950	990	2.994
397	2.599	463	2.666	529	2.723	597	2.776	663	2.821	729	2.863	797	2.901	894	2.951	993	2.996
399	2.601	465	2.667	531	2.725	599	2.777	665	2.823	731	2.864	799	2.903	897	2.953	996	2.997
				533	2.727					733	2.865					1,000	3.000

CHAPTER VIII.

ACCURACY.

INTRODUCTORY.

THERE is not in existence a perfectly accurate measurement, physical or economical, just as there is no perfectly straight line or perfect fluid. We can best illustrate the nature of economic measurements by considering that of physical. It is easy to weigh substances accurately to 1 gram : then by obtaining a good balance, we can, as our apparatus is improved, weigh accurately to a centigram, milligram, and one-tenth of a milligram; but for accuracy beyond this the balance fails us. Similarly in measuring angles, the naked eye can distinguish an object which subtends one-thirtieth of a degree; with a sextant a measurement can be taken correctly to fifteen seconds of arc; the Greenwich astronomers can make observations correct to one-hundredth part of a second, but we again come to a point beyond which precision is unattainable.

The nature of measurement.

In such cases the result is stated as correct to a milligram, or whatever it may be; in the same way we speak of an estimated sum of money correct to a pound.

A task which has considerable resemblance to some statistical estimates, is the measurement of the parallax of the sun, which determines its distance from the earth. During the eighteenth century astronomers estimated it as 10″, equivalent to 96,000,000 miles. As methods of observation and instruments were improved, observers began to agree that the whole number of seconds was 8, but gave various estimates for the first decimal figure. Since 1865 there have been very few estimates which have not given 8 as the nearest figure for this place (8·8″), while more recent observations agree in making the parallax from 8·76″ to 8·78″. We may, therefore, consider that the distance is now accurately

Physical and statistical measurements.

178

known to within 1 in 400. Notice in this connection, first, that the earlier observations have been subject to corrections; secondly, that better agreement has been attained as time has gone on; thirdly, that neither absolute agreement nor absolute accuracy have yet been obtained. So it is with statistical measurements; we might instance the gradual settlement of the curve representing expectation of life, the measurement of the fall in prices, and the development of wage statistics.

Again in physical measurements, though we can sometimes reach a very high degree of accuracy, as, for instance, in the weight of a cubic foot of water which could doubt- Degrees of possible accuracy. less be known correctly to one part in a million, in other cases we are glad if we can measure to one part in ten, as, for instance, in the distance of the nearest fixed star from us, which is, roughly, from 34 to 37 billion miles. So in statistics it is something if we know that the total capital of the United Kingdom was between $7\frac{1}{2}$ and 10 thousand million pounds in 1885, or if we know that the average weekly wage of working-men in full work was from 21s. to 27s. in 1886. The weak point in such statements is that often when we have made an estimate, which we know to be inexact, we are not able to give any estimate of the limits of the error. We are not so definite as *The Modern Traveller* who

> ". . . knew the weather to a T,
> The longitude to a degree,
> The latitude exactly."

We are not able to say " our estimate is 24s. 5d.; this is probably correct within 3d., and it is not possible that we are as much as 6d. wrong "; whereas in physical measurements we can often give the result as correct to the smallest graduation of the instrument employed.

On the other hand, though we cannot obtain exactness, we can in many cases estimate to that degree of accuracy which is required for practical purpose. In common use The accuracy only a certain conventional accuracy is needed. generally needed. Thus, to take some miscellaneous instances, the area of an estate is given in acres, roods, and poles, but not correct to square yards; the market prices of shares do not change less than $\frac{1}{16}$; we keep the day, not the hour, of our birth; railway time-tables do not show seconds; ocean steamers are timed to start at certain hours, not minutes; height is measured correct

to one-tenth of an inch; a hundred yards race is timed to one-tenth of a second. Similarly in statistical estimates, we seldom need that our results shall be accurate within one per thousand, or even 1 per cent. One per thousand of the working week is less than three minutes; 1 per cent. of the week's wage is only 6d. We do not care to know the population of London within 100, the expenditure of the Exchequer within £1000, or the expectation of life within a day. It is often possible to attain practical accuracy within such limits.

DEFINITION OF ERROR.—For purposes of measurement we may take the following definition :—*The relative error in an estimate is the ratio of the difference between the estimate and the true value, to the estimate :* the error is to be reckoned positive when the true value exceeds the estimate.

Thus if the average weekly wage of agricultural labourers was in reality 14s., and we estimated it as 13s., our error would be $\dfrac{14 - 13}{13} = \dfrac{1}{13}$, or 7·7 per cent.; if we had estimated it as 15s., the error would be $\dfrac{14 - 15}{15} = -\dfrac{1}{15}$, or − 6·6 per cent.

In algebraic notation, if u be the measurement of a quantity whose true value is u^1, then $\dfrac{u^1 - u}{u}$ is the error in the estimate, which we shall call e; so that $e = \dfrac{u^1 - u}{u}$, and $u^1 = u\,(1 + e)$.* e thus defined is the *relative* error, while ue is the *absolute* error.

In the nature of things, when we are dealing with errors, we do not know their magnitude; the most we can know Statement of errors. is their probable and possible extent. We might estimate, for instance, the percentage of unemployed in a certain year as 4·5, and add, from information in our possession (coming from a study of wage-bills or the reports of relief agencies), that we considered this to be within ·5 of the fact; we should then write the number 4·5 ± ·5, meaning that the error in the estimate as defined above was unlikely to be more than $\dfrac{·5}{4·5} = \dfrac{1}{9}$, or 11 per cent., the corresponding absolute error being ·5. In such a case we can also

---

* It is sometimes more convenient to write $u = u^1\,(1 + \epsilon)$, reckoning the error relatively to the true value. Then $\epsilon = -e + e^2$ approximately, and when e is less than 10 per cent. we may take $\epsilon = -e$.

give definite limits. The percentage unemployed must lie between 0 and 100; and if we could actually enumerate 1 per cent. of the working-class as out of work, and also 92 per cent. as in work, we should know that the number required was between 1·0 and 8·0 per cent., and the maximum error in our estimate, 4·5, was $\frac{3·5}{4·5} = \frac{7}{9}$, or 78 per cent. Even this is more precise than the original statement, " the percentage is 4·5, error unknown." By further investigation we might perhaps bring the limits of error nearer to each other, and decide that it was practically certain that the percentage required was between 3·5 and 4·5; then we ought to say " the number unemployed is ·04 . . . of the working-class, the estimate being correct to the last figure given." This statement is of the same nature as, " The body weighs 15 lbs. 3 oz., correct to an ounce."

While, on the one hand, it is clear that we cannot often obtain close definite limits to our errors, on the other we can very often see that some of the digits in a total are almost certainly right and others almost certainly wrong. Thus when we see in the Registrar-General's Report that the population of the United Kingdom in 1895 was 39,124,496, the estimate being made from the census of 1891, and the increase calculated on the basis of the increase since 1881, we may be certain that the last two, or the last three, digits are no better than guess-work; while the first two, or the first three, are correct. Thus the statement should read: Population was 39·1 millions, or 39,124,000 ± 5000, or whatever figures our examination of the varying rate of progress of the population led us to adopt, and such a statement is actually more correct than the previous one.

It is the custom in many classes of estimates to give the figures to the uttermost farthing. This is possibly right in official publications; for the duty of the office   Neglect is to receive and tabulate returns, stating how   of minutiæ. and whence they came, and it may leave to the economist or the statistician the task of deciding the degree of accuracy pertaining to them. But in summary descriptions and accounts, and in scientific estimates, it is not merely unnecessary to give these last figures (both because they are not accurately known, and because they generally have no impor-

tance to the argument or significance to the reader), but it is positively inaccurate. The easiest way to avoid the inaccuracy is simply to state totals in so many thousands (*e.g.*, the earth is 8000 miles in diameter), or if for any reason more exact measure be required (as when we are comparing the equatorial diameter with the smaller one through the poles), the scientific way is to give the number as far as it has been fairly calculated, and to indicate its precision.

RULES FOR COMPUTING THE EFFECT OF RELATIVE ERRORS.

We may now give some rules connecting the errors of a complex estimate with those of the elements which form it.

I. *The error in an estimated sum is equal to the sum of the errors in the parts when each is multiplied by the ratio of the corresponding part to the sum.*

For if we estimate n quantities as $u_1, u_2 \ldots u_n$, and their sum as u, so that $u = u_1 + u_2 \ldots u_n$, and the errors of the quantities are $e_1, e_2 \ldots e_n$, and that of the sum is e :

Error in sum.

then the true value of the sum is $u\,(1+e)$, and the true values of the parts are $u_1\,(1+e_1)$, $u_2\,(1+e_2) \ldots$, so that—

$$u\,(1+e) = u_1\,(1+e_1) + u_2\,(1+e_2) + +,$$
$$\text{but} \quad u = u_1 \qquad\quad + u_2 \qquad\quad + + ;$$
$$\text{hence, by subtraction,} \quad ue = u_1\,e_1 \quad + u_2\,e_2 \quad + +,$$
$$\text{and} \quad e = e_1 \times \frac{u_1}{u} \quad + e_2 \times \frac{u_2}{u} \quad + +.$$

The formula is easily adapted to the case where some of the parts are subtractive.

To take an arithmetical example, if average working-class expenditure on food, clothes and rent was estimated in 1914 as 25s., 5s. 6d., and 6s. 6d. respectively, while the true averages were 27s., 4s. 6d., and 6s., so that the errors are $+\frac{2}{25}, -\frac{2}{11}$, and $-\frac{1}{13}$, then the error in the sum of the three is—

$$+\frac{2}{25} \text{ of } \frac{25}{37} - \frac{2}{11} \text{ of } \frac{5 \cdot 5}{37} - \frac{1}{13} \text{ of } \frac{6 \cdot 5}{37} = + \cdot054 - \cdot027 - \cdot0135$$
$$= + \cdot0135 \text{ or } + 1\frac{1}{3} \text{ per cent.}$$

We can apply the rule to the important case where we can estimate a great part of a required total with considerable

accuracy, while we are ignorant of a smaller part. Thus we may receive returns from several unions that 33,650 are out of work, and have reason to know that the error is not more than 1 per cent., while some smaller unions do not send any returns; we make an estimate for the smaller unions, say that 1000 of their members are unemployed, and suppose a very large error, say $\frac{2}{3}$ or 67 per cent. Then the error in the total is less than—

$$\frac{1}{100} \text{ of } \frac{33650}{34650} + \frac{2}{3} \text{ of } \frac{1000}{34650} = \cdot 029 \text{ or less than 3 per cent.,}$$

an error very much nearer that of the larger returns than that of the smaller. In the preceding sentence we say " less than," because we assume that we have taken an outside limit for the smaller errors.

II. *The error in the arithmetic average of several estimates is the sum of the errors of these estimates, when each is multiplied by the ratio of the corresponding estimate to that of the sum of the estimates.*

For if $m_1, m_2, \ldots m_n$ are n estimates of quantities whose true values are $m_1 (1+e_1), m_2 (1+e_2), \ldots$, the estimated and   Error in true averages are respectively—   average.

$$\frac{m_1+m_2+ \ldots m_n}{n} \text{ and } \frac{m_1 (1+e_1)+m_2 (1+e_2)+ \ldots +m_n (1+e_n)}{n}$$

and the error in the average is—

$$\frac{\dfrac{m_1 (1+e_1)+m_2 (1+e_2)+ +}{n} - \dfrac{m_1+m_2+ +}{n}}{\dfrac{m_1+m_2+ +}{n}} = \frac{e_1 m_1+e_2 m_2+ +}{m_1+m_2+ +}$$

$$= e_1 \times \frac{m_1}{S \cdot m} + e_2 \times \frac{m_2}{S \cdot m} + \ldots$$

where S denotes the sum of all the m's.

It is easily seen that no individual error can have much influence on the result, that the error in the average would be nearly of the same magnitude as one of the individual errors, if these were not very unequal and all positive or all negative, and that if, as is generally the case, some are positive and some negative (a point we shall consider presently), the error would be considerably lessened.

III. *The error in a weighted average is the sum of* (1) *an error due to errors in the quantities, similar to the error of an unweighted average, and* (2) *an error due to errors in the weights, which becomes very small when the original quantities are nearly equal.*

Let $W_1, W_2 \ldots W_n$ be estimated weights applied to n estimated

Error in weighted average. quantities $M_1, M_2 \ldots M_n$, and let the true values of the weights be $W_1 (1+\epsilon_1), W_2 (1+\epsilon_2) \ldots$ and of the quantities be $M_1 (1+e_1), M_2 (1+e_2), \ldots$

Write $M_w = \dfrac{SWM}{SW}$, so that M_w is the estimated weighted average, and let $M_w (1+E)$ be its true value.

Then
$$M_w . E = \frac{SW (1+\epsilon) M (1+e)}{SW (1+\epsilon)} - \frac{SWM}{SW}$$

$= [SW.S\{W_\iota M_\iota (1+\epsilon_\iota) (1+e_\iota)\} - SWM.S\{W_\iota (1+\epsilon_\iota)\}] \div SW.SW (1+\epsilon)$,
where the suffix ι denotes any selected quantity, etc.

Then—

$E.SWM.SW (1+\epsilon) = SW.SW_\iota M_\iota e_\iota + SW.SW_\iota M_\iota (\epsilon_\iota + e_\iota \epsilon_\iota) - SWM.SW_\iota \epsilon_\iota$.

Now suppose E, e_ι, ϵ_ι to be as small as $\cdot 1$, and neglect products which are as small as $\cdot 01$.

$$E . SWM . SW = SW . SW_\iota M_\iota e_\iota + S\{W_\iota (M_\iota . SW - SWM) \epsilon_\iota\}$$
$$\therefore E = \frac{SW_\iota M_\iota e_\iota}{SWM} + \frac{S\{W_\iota (M_\iota . SW - SWM) \epsilon_\iota\}}{SWM . SW} .$$

The term involving e_ι, the error in a quantity, is the same as that in Rule II., if $W_1 M_1$ is written for m_1, etc.

The coefficient of ϵ_ι needs further analysis.

Since $SWM = M_w . SW$, $M_\iota SW - SWM = SW . (M_\iota - M_w) = m^1_\iota SW$, where m^1_ι is the excess of a quantity over the weighted average.

$$\therefore E = S \frac{W_\iota M_\iota}{SWM} . e_\iota + S \frac{W_\iota m^1_\iota}{SWM} \epsilon_\iota .$$

Hence the resulting error due to the errors in *quantities* involves the magnitudes M_1, M_2, etc., while that due to the errors in *weights* involves only the deviations of these quantities from their weighted average. These deviations are individually small if the dispersion of the quantities about their mean is small relatively to that mean. Further, the sum of the coefficients $W_\iota m^1_\iota = SW_\iota M_\iota - M_w SW = 0$; if the errors in weights are all equal the resulting error in the average is zero, as is evident *a priori*, and if positive errors are not generally found with positive deviations (m^1_ι) and negative with

negative, and if large errors are not generally found with large weights (and vice versa), the sum of the terms $W m^1 \ell_\ell$ tends to be small.

• Hence the errors in weights have an effect which not only diminishes from the same causes as affect the errors in quantities, but also have coefficients which have a strong tendency to neutralise one another, unless the magnitudes of the errors, quantities and weights are associated with each other in special ways. Great errors are required in the weights, if many quantities are involved, to make an appreciable error in the average. In fact, *the errors in quantities have so much more influence than those in weights, when once the weights have been reasonably estimated, if the quantities are not very unequal, that errors in the weights can very frequently be neglected.* Several numerical examples of this principle were given in the section on weighted averages.

IV. *The error in a product is approximately the sum of the errors in its factors, due regard being paid to sign.*

For if $f_1, f_2, \ldots f_n$ are the estimated factors, whose true values are $f_1(1+e_1), f_2(1+e_2), \ldots$, then the error of the product

$$= \frac{f_1(1+e_1) \cdot f_2(1+e_2) \cdot \ldots \cdot - f_1 \cdot f_2 \cdot \ldots}{f_1 \cdot f_2 \cdot \ldots} \qquad \text{Error in product.}$$

$= (1+e_1) \cdot (1+e_2) \ldots - 1 = e_1 + e_2 + \cdots + e_n$, if we neglect products of two or more e's.

The e's are equally likely, *a priori*, to be positive or negative. If two e's are of different signs, they tend to neutralise one another. The error in a product may be great if all the errors of the factors are of the same sign, even if they are small individually.

For example, if we estimate that 100 men are earning on the average 25s. each, while in reality there are 105 men earning 26s., the error in the estimated total sum earned is, by formula,

$$\frac{5}{100} + \frac{1}{25} = \cdot 09.$$

If, with the same estimates, the real quantities had been 105 and 24s., the error in the product would have been

$$\frac{5}{100} - \frac{1}{25} = \cdot 01.$$

V. *The error in a ratio is approximately the difference between the errors in its two terms, due regard being had to sign.*

For if u_1, u_2 be the estimated terms, whose true values are
Error in ratio. $u_1 (1+e_1)$ and $u_2 (1+e_2)$, then the error in the ratio is—

$$\frac{\dfrac{u_1 (1+e_1)}{u_2 (1+e_2)} - \dfrac{u_1}{u_2}}{\dfrac{u_1}{u_2}} = \frac{1+e_1}{1+e_2} - 1 = \frac{e_1-e_2}{1+e_2}$$

$$= (e_1-e_2)(1-e_2+e_2{}^2-e_2{}^3+ -)$$
$$= e_1-e_2, \text{ if we neglect terms of the}$$

second order in the e's.

If the errors in the terms are both positive or both negative, they tend to neutralise one another ; if they are also nearly equal, the error in the ratio becomes very small.

We can apply Rule V. to the error in comparison of two averages of similar quantities estimated at different dates.

With the same notation as under Rules II. and III., using m, e, ϵ, for the quantities at one date, and m^1, e^1, ϵ^1, for similar quantities at another date, then the error in the ratio of the simple average of $m_1{}^1$, $m_2{}^1$ . . . to the simple average of m_1, m_2 . . . is—

$$S\left\{e^1\left(\frac{m^1}{Sm^1}\right)\right\} - S\left\{e\left(\frac{m}{Sm}\right)\right\}$$

$$= \left(e_1{}^1 \cdot \frac{m_1{}^1}{Sm^1} - e_1 \cdot \frac{m_1}{Sm}\right) + \left(e_2{}^1 \cdot \frac{m_2{}^1}{Sm^1} - e_2 \frac{m_2}{Sm}\right) + +$$

Now if the quantities have not changed much during the period between two observations, the fraction $\dfrac{m^1}{Sm^1}$ will differ little from $\dfrac{m}{Sm}$, and so on.

Neglecting these differences in comparison with the quantities themselves, a legitimate process when we are estimating the approximate influence of errors, we have—

Error in the ratio of the simple averages $= S\left\{\dfrac{m_1{}^1}{Sm^1}(e_1{}^1-e_1)\right\}$

If the two estimates have been made under nearly similar circumstances, leading to similar chances of errors, $e_1{}^1$ and e_1 are likely to be not only of the same sign, but nearly equal.

Write d_1, d_2 . . . for $(e_1{}^1-e_1)$, $(e_2{}^1-e_2)$ . . . , and we have—

Error $= S . \left\{d_1 . \left(\dfrac{m_1{}^1}{Sm^1}\right)\right\}$, where the d's may be small.

The corresponding analysis for the error in the ratio of two weighted averages is too complicated to be given here ; * but

---

* It will be found in the *Statistical Journal*, 1911, pp. 85 *seq.*, and in a modified form in Part II, Appendix, Note 7.

using the principle that errors in weight are less important than errors in quantity, which applies with slight modifications, we may use the formula just given for the first approximation to the error in the ratio of two weighted averages. This formula may be put in words :—

VI. *The error in the ratio of two averages of similar series of quantities, estimated at different dates, is approximately equal to the sum of the differences between the errors in the corresponding terms of the two series, each multiplied by the ratio of the latter of these corresponding terms to the sum of all the terms at the latter date.*

This rule is so important that it will be worth while to illustrate it by an example, in which a further quantity will be introduced.

If in each of two years we are able to estimate, as in our example under Rule I., one part of a total more accurately than another part, we can use the following formulæ :—

	First Year.	Second Year.
Estimated numbers or weights	w; error ϵ;	w^1; error ϵ^1
Estimated average income, or quantity	m_1; error e_1;	m_1^1; error e_1^1
Estimated number, less accurately known	rw; error in r, ρ;	$r^1 w^1$; error in r^1, ρ^1
Estimated income	m_2; error e_2;	m_2^1; error e_2^1

e_1 and e_1^1 are, by hypothesis, less than e_2 and e_2^1.

Error in average for first year—

$$\frac{\dfrac{w(1+\epsilon).m_1(1+e_1)+r(1+\rho).w(1+\epsilon).m_2(1+e_2)}{w(1+\epsilon)+r(1+\rho)w(1+\epsilon)} - \dfrac{wm_1+rwm_2}{w+rw}}{\dfrac{wm_1+rwm_2}{w+rw}}$$

$$=e_1\frac{m_1}{m_1+rm_2}+e_2\frac{rm_2}{m_1+rm_2}+\rho\frac{r}{1+r}\cdot\frac{m_2-m_1}{m_1+rm_2}$$

if we neglect products of e and ρ.

Here the errors, e_2 and ρ, connected with the less accurately known part, are each multiplied by r, the ratio of the weight of that part to the weight of the better known part, ρ is multiplied by m_2-m_1, which in many cases is small, while e_1, the remaining error, is by hypothesis small.

If for simplicity of argument we assume that the ratio of the unknown part to the whole (but not the error in estimating it)

has remained unchanged, and also that the ratio of the estimated average incomes of the two parts has not altered, we have for the error in comparison—

$$(e_1{}^1-e_1) \cdot \frac{m_1}{m_1+rm_2} + (e_2{}^1-e_2) \cdot \frac{rm_2}{m_1+rm_2} + (\rho^1-\rho)\frac{r}{1+r} \cdot \frac{m_2-m_1{}^1}{m_1+rm_2}$$

Thus in estimating the change in average wages of Scotch agricultural labourers, we have figures similar in character to the following :—

1867. MARRIED PLOUGHMEN.				1892. CORRESPONDING NUMBERS.		
Estimated number -	1,000	Average income,	£36	1,200	£49	0 0
Supposed true number	1,010	,, ,,	35	1,220	48	0 0

FARM-SERVANTS.						
Estimated number -	200	Average income—		240		
		Money -	£21		£27	5 0
		Estimated value of board -	13		14	0 0
		Total -	£34		£41	5 0
Supposed true number	220	Total income -	£37	240	£47	0 0

Here $w=1,000$, $m_1=36$, $r=\frac{1}{5}$, $m_2=34$, $w^1=1,200$, $m_1{}^1=49$, $r^1=\frac{1}{5}$, $m_2{}^1=41\frac{1}{4}$, $\epsilon=\frac{1}{100}$, $e_1=-\frac{1}{36}$, $\rho=\frac{9}{101}$, $e_2=\frac{3}{34}$, $\epsilon^1=\frac{1}{60}$, $e_1{}^1=-\frac{1}{49}$, $\rho^1=-\frac{1}{61}$, $e_2{}^1=\frac{23}{165}$.

Here it is supposed that we have overvalued the income of the married ploughmen, and undervalued that of the farm-servants in both cases. We suppose, as is the fact, that the value of the board and other perquisites of the farm-servants cannot be estimated with precision, and that the proportionate numbers in the two classes are not accurately known.

Substituting in the above formula we find that the error in the estimated ratio of the average incomes of the two classes together in the two years is—

+ ·0062, due to errors in estimates of income of ploughmen.
+ ·0081, ,, ,, ,, servants.
+ ·0008, ,, ,, ratios of the numbers in the two classes.

Thus the last error, due to weights, is very small, and the second error, due to ignorance of the value of board, is reduced by the smallness of the number employed to a magnitude comparable with the first.

The whole error is, therefore, by formula + ·0151. Going

to the actual figures, we find the estimated ratio of the second to the first to be 1·3376 to 1, and the supposed true ratio to be 1·3529 to 1; that is, the error is $\dfrac{·0153}{1·3376} = + ·011.$

The difference between the two methods of calculation is accounted for by the neglect of the less important terms.

It is to be noticed that the error in the ratio of two quantities is not the same as the error which we might be inclined to estimate, the error in the percentage increase. Thus in the case just taken, the estimated and true percentage increases are 33·8 and 35·3, and the relative error in the percentage increase is ·045. For accuracy in such calculations, then, we require the error found by formula, according to Rule VI., to be very small.

Another example is found from the well-known difficulty of estimating the relative importance of expenditure on clothing in a workman's family budget.

The following estimates were used in the Report on the Cost of Living, 1918 (Cd. 8980, pp. 7, 18 and 23).

SKILLED WORKMEN, AVERAGE WEEKLY EXPENDITURE.

	1914.	1918.	Ratio.
Food - - -	27s.	49s. 10d.	1·84
Clothing - - -	7s.	13s. 9d.	1·96
Together - - -	34s.	63s. 7d.	1·864

Here we take $w=27$, $r=\frac{7}{27}$, $m_1=1·84$, $m_2=1·96$.

Suppose that r ought to have been taken as $\frac{1}{4}$, and m_1, m_2 as 1·90, 2·10.

Then $e_1=\frac{3}{92}=·0326$, $e_2=\frac{1}{14}=·0714$, and $\rho=\frac{2}{7}=·286$.

The resulting error by formula is—

+ ·0256, due to error in the ratio of food expenditure at the two dates.
+ ·0155, ,, ,, ,, clothing ,, ,, ,,
+ ·0030, ,, ,, ratios of the expenditures on clothing and food.

And the whole relative error is ·044.

The effects of the errors are in the reverse order of their magnitude, and the great error in the clothing ratio barely affects the second decimal place in the result.

If, however, m_2-m_1 had been larger, that is if the estimated

increase in expenditure on clothing had been much greater than that on food, the effect of this error would have been proportionately more.

We return to the whole question of relative errors as illuminated by the theory of probability in Part II, Chapter IV, below.

BIASSED AND UNBIASSED ERRORS.

In the consideration of all errors in averaging or comparing, it is important to distinguish two classes of errors, those which

Errors are biassed or un-biassed.

are biassed and those which are unbiassed. The difference can be made clear by illustrations. If a number of men are sent to investigate the condition of an industry in different places, with a view of proving that wages are high, conditions of work healthy, and so on, they would probably, by examining only the best conducted works, and taking the wages only of the more skilled and regular workmen, produce an average for each town which would be too high. On the other hand, if there was no brief to be held, but the investigation was impartial, the commissioners would in some towns take too high an average, in others too low, according to their idiosyncrasies and to circumstances. In the first case, the errors would be biassed, all in the same direction, all tending to increase the average, whose error would be equal to the average error in the different towns. In the second case, the errors would be unbiassed, just as likely to be in excess or defect, and the more estimates made, the smaller would the resulting error be. The following figures would illustrate this :—

	Fact.	Biassed Estimate.	Unbiassed Estimate.
	s.	*s.*	*s.*
Average Wages in District—*a*	24	25	24
,, ,, *b*	23	25	25
,, ,, *c*	26	27	25
,, ,, *d*	27	28	28
,, ,, *e*	28	30	27
Averages - - - -	25.6	27	25.8
Errors - - - - -	...	5.2 %	1 %

In measuring the distance of a bicycle ride on a mile-stoned road, it is found that the distances between successive milestones are not exact, but perhaps 50 to 100 yards out; but it may be nearly as likely that the errors will be in excess or defect, and the greater the distance gone the smaller will be the error, as defined. The errors are unbiassed. If, on the other hand, the bicyclist trusts to his cyclometer, he will have to deal with a biassed error, for the instrument will not fit the wheel exactly, but will always register say 1800 yards when the machine has gone a mile. This is a case where the bias can be measured and allowed for, whereas the unbiassed errors must be left to eliminate themselves. It is frequently the case that biassed errors are due to a wrongly graduated instrument; unbiassed to separate faulty measurements.

In the census returns, the fact that many women return themselves as younger than their birth certificate states, causes a biassed error in the average age of the population; the fact that people frequently return their ages at the nearest round number causes unbiassed error, and on the whole affects the average little. It is not improbable that in the Wage Census of 1906, there was some tendency to obtain returns from the more liberally conducted establishments in some industries; this causes a biassed error in the average obtained. With these illustrations we can pass on to another principle of great importance. Unbiassed errors are of little importance compared with biassed errors in a simple estimate; but biassed errors diminish when the ratio of two similar estimates is taken.

Relative importance of biassed and unbiassed errors.

For in an average of several quantities, which have biassed errors (η_1, η_2 . . .) and unbiassed errors (e_1, e_2 . . .), it is easy to see from Rule II. that the resulting error may be written

$$S\left(e_1 \cdot \frac{m_1}{S \cdot m}\right) + S\left(\eta \cdot \frac{m_1}{Sm}\right).$$

In the first term, the errors being unbiassed, many of them are positive, many of them negative, and they tend to neutralise one another; in fact, if E is typical of the errors e_1, e_2 . . . e_n, then a first approximation to the error arising from them in the average is

$$\frac{2E*}{3\sqrt{n}}.$$

* It is as likely as not that so great an error would be obtained. See Part II, Chap. IV.

Thus in the average of one hundred measurements, whose individual unbiassed errors are about $\frac{1}{10}$, the resulting error may be no greater than $\frac{2}{3} \times \frac{1}{10} \div \sqrt{100} = \frac{1}{150}$. There is no counterbalancing tendency, on the other hand, in the biassed errors; if each estimate was 10 per cent. in excess, then the average is also 10 per cent. in excess. When aiming at Great effect of accuracy our principle always is to take care of biassed errors. the pounds, and let the pence take care of themselves; and it is quite futile to diminish the unbiassed errors, that is to increase the precision of our measurements, while a large biassed error runs through them all. If we do not know of the existence of biassed errors, which in reality pervade our estimates, there is no remedy; if we do know of them, we are likely to obtain more accuracy by the most erroneous corrections for them than by neglecting them; for when we make unbiassed corrections for our biassed errors, we reduce them to unbiassed errors, and then the more terms we include in our average the smaller is our resulting error. If, for instance, we find that the average weekly wage of agricultural labourers throughout the country is 13s., and by considering the circumstances of the thousand returns which we may suppose led to this average we have reason to suppose that an error of 1s. would be typical of the unbiassed errors in them, then an error of $\frac{2}{3}$ of $\frac{1s.}{\sqrt{1000}}$, that is only a farthing, may be expected to result in the average. We have here a totally illusive accuracy; the part of the labourer's income which we have not included, payments at haytime and harvest, facilities for piece-work, cheap rent for cottage and land and smaller perquisites, is not capable of exact calculation. If we omit all these entirely we shall leave an error in our average of 2s. or so; but we make *individual* estimates of these additions, in all the thousand cases, though each estimate may be 2s. wrong, if there is no bias, the resulting error in the average may be expected to be $\frac{2}{3}$ of $\frac{2s.}{\sqrt{1000}}$, that is only $\frac{1}{2}d.$: our whole error now may be less than 1d., instead of 2s. In estimating the accuracy of published averages, these principles should be always borne in mind, and the possibility of biassed errors always considered.

When we are dealing with the errors of a ratio the case is quite different. The error of a ratio is approximately equal to the difference between the errors in its terms; if $\quad$ Accuracy of η, η^1 and e, e^1 are the biassed and unbiassed errors $\quad$ comparisons. in the terms, then by Rule V. $(\eta^1 - \eta) + (e^1 - e)$ is the error in the ratio. Now the unbiassed error $(e^1 - e)$ is likely to be of nearly the same magnitude as either e or e^1 ; * if, as in the above example, e and e^1 are unlikely to be much greater than $\dfrac{2}{5}$, $(e^1 - e)$ would be unlikely to be much greater than $\dfrac{3}{5}$. But $(\eta^1 - \eta)$, the result of the biassed errors, will, if the bias in both terms of the ratio was in the same sense (positive in both, or negative in both), be less than the original errors. If we have made the estimates of both terms on precisely similar methods, if we have asked the same questions of the same classes of persons, included and omitted the same details on both occasions, we shall have made nearly the same errors of bias in both estimates. To return to our previous illustration, if we have made the glaring mistake of omitting everything except average weekly wages in the income of an agricultural labourer on both occasions, the only resulting error in the ratio will be that due to the *change* in the proportion that these extra payments bear to ordinary wages, which in short periods is likely to be small. Or, if we had taken summer wages as the average for the year in both cases, the error in the ratio will depend only on the *change* in the relation of summer wages to that average. Hence the error in the ratio of two estimates at different dates of a slowly changing quantity is, if the estimates are made on similar methods, often much smaller than the error in either estimate singly; for the unbiassed error is little greater, and the more important biassed error is much diminished. We need not now know of the existence of the biassed errors; they will disappear of themselves. If we are aware that there are biassed errors, and have any means of making fairly good estimates of them, it will be worth doing; but we shall make a great mistake if we correct the bias in one year and leave it uncorrected in another. For purposes of comparison it is very seldom of much use and often of great

---

* If E is the probable error in e or e^1, then $E . \sqrt{2}$ is the probable error in their difference; see Part II, Chap. III.

disutility to make the later estimate more accurate than the earlier. The error resulting from unbiassed errors can indeed be diminished a little,[*] but the error resulting from the more important biassed errors will only be increased. All Government officials and others who compile annual returns are in a dilemma : to make their annual statements accurate in themselves, they should always be straining after improvements, they should always be watching for changes in the quantities measured and adapting their methods and tabulations to these changes; but to make their annual returns comparable with each other, they should be absolutely conservative, and cling to any mistakes they or their predecessors have made in the past with all the strength red tape can give them, being careful, however, not to add to the mistakes or make new omissions. The dilemma can in some cases be avoided; for when an improved method is introduced, the tabulation can sometimes be given for a few years both on the old and on the new plans; then when the difference introduced by the change is known, the earlier figures can be brought to the greater precision of the later. Thus the Board of Trade since 1898 has included in the tabulation of exports ships which, leaving our shores with merchandise, are themselves sold to a foreign owner; and we have the following tabulation :—

Need for uniformity in structure of serial returns.

	1899.	1898.
Exports of Home Products (exclusive of ships sold to foreigners) - - -	£255,465,000	£233,359,000
Re-exports of Home and Colonial Merchandise	65,020,000	60,655,000
Total - -	£320,485,000	£294,014,000
Value of New Ships exported	9,195,000	Not stated.
New total -	£329,680,000	

---

* For if E and E_1 be typical of the unbiassed errors at the two dates, then $\sqrt{E_1{}^2+E^2}$ is typical of the error in the ratio, which diminishes with either E or E_1. See Part II, Chap. IV, formula (66).

Ignorance of slight alterations in the collection and tabulation of material has been the cause of many statistical mistakes.

To sum up the chief results of this chapter : there are two processes which tend to accuracy—*averaging*, which diminishes unbiassed errors; and *comparison*, which dimin- **Results.** ishes biassed error. *The errors in weights are seldom so important as the other errors which are present in estimates.* Errors in a result cannot, of course, be calculated, but can be expressed in terms of errors in the items, from which it comes; we cannot attain certainty, but we can indicate processes which diminish errors, and with the help of mathematics measure the extent of diminution. Initial errors are diminished most, when we calculate the *ratios of weighted averages of similar and similarly estimated quantities.* Index-numbers, which we discuss in the next chapter, are examples of this class.

The accuracy resulting from the process of sampling requires more mathematical treatment, and is dealt with in Part II, Chapters II, and IV.

CHAPTER IX.

INDEX-NUMBERS.

THE discussion of index-numbers supplies so good an illustration of the principles laid down in the last chapter, and index-numbers are so important in themselves, that, though it is our intention to avoid special questions, it will be worth while to devote a chapter to them.

Index-numbers are used to measure the change in some quantity which we cannot observe directly, which we know to **Function of** have a definite influence on many other quantities **index-numbers.** which we can so observe, tending to increase all, or diminish all, while this influence is concealed by the action of many causes affecting the separate quantities in various ways. Thus, to take three of the quantities to which index-numbers are applied, the change in the relation of the precious metals to the work to be done by them affects prices of all commodities, but very many other causes are at work affecting the prices of separate groups of commodities; there are general causes tending to raise the wage of a week's work of average skill, but this general increase is concealed by numberless minor causes affecting different grades of labour in different degrees; the change in the consumption of goods by the working or other classes is a sufficiently definite quantity, but it can only b‑ measured indirectly by observing the varying changes in the consumption of individual articles.

The use of index-numbers is not, however, confined to these instances, but is nearly co-extensive with the field of statistics; for we have limited the term statistics to the measurement of complex groups and their changes; the object of statistics is to measure the action of the general laws which govern a heterogeneous group, and the changes produced by general forces can be measured, as a rule, only by their effect in individual cases; thus the method of index-numbers is at once applicable to the

disentanglement of that which is common to the whole group from those variations which are special to individual items.

In the more restricted sense a series of index-numbers is a series of weighted averages, calculated periodically, where the quantities averaged are similar (prices or wages), and the weights are defined so as to give the actual average of the whole group concerned in each measurement. In its less restricted sense a *series of index-numbers is a series which reflects in its trend and fluctuations the movements of some quantity to which it is related.* Where the weights and the quantities are both known exactly, the method of index-numbers is merely a convenient way of expressing straightforward arithmetical results in a simple manner; this simplicity can be nearly realised in index-numbers of prices of exports. Where the quantities are samples selected from a wide group, and there is no obvious method of deciding their relative importance, the index-numbers have a less direct relation to the movement of a definable and measurable phenomenon; such is the nature of most price index-numbers and of some wage index-numbers. Where the quantities are not direct measurements of examples of the phenomena which it is desired to study, but of allied phenomena, then the connection between the series of index-numbers and the phenomena is indirect; such, in fact, are most of the index-numbers of wages and of employment.*

The most ordinary way of forming an index-number is intermediate between the extremes of exactness and of indirect relation. Thus in the Labour Department's index-number of the change of rates of wages, the objective is presumably to find numbers, year by year, whose ratios are the same as the ratios of the average rates of weekly wages of persons in regular industrial work in the United Kingdom; at least, the numbers are generally quoted in this sense, and the heading in the Abstract of Labour Statistics is " General Course of Wages in the United Kingdom " (*e.g.*, XVIth Abstract, Cd. 7131, p. 82). This index-number is obtained by selecting some hundreds of recognised time or piece rates, expressing each as a percentage of its amount in 1900, and averaging the results year by year. The choice of weights in this average is indirect; each of the five groups (building, coal-mining, engineering, textiles, agricul-

---

* Parts of these pages are taken from the *Statistical Journal*, 1912, pp. 791-5.

tural) is taken as of the same importance, while the building group contains 74 items, agriculture 115, etc. Mr. Sauerbeck obtains his index-number of prices by selecting the prices of typical commodities, and weights them by the device of duplicating quotations for the more important. Thus in these and other cases we have a selection of " quantities," whether deliberately or by accident, and an assignment of " weights," whether directly or indirectly. It is then hoped that' the numbers will move in direct proportion to the phenomenon, average wage or average level of prices, whose measurement is attempted.

In such cases three points call for consideration—(1) The nature and extent of the group and the nature of its special property whose general change is studied. (2) The method of choosing samples. (3) The effect of weights. (1) With Mr. Sauerbeck's numbers the group is Prices of wholesale commodities in the United Kingdom ; and with other index-numbers the groups are the prices of goods exported, of goods imported, and so on. In the Labour Department's wage index the group is two-fold and consists of (a) rates of weekly time wages, (b) piece rates, and the result is hybrid. It is essential to define both the extent of the group and the property or attribute which is to be measured. The property is sometimes elusive, as the " purchasing power of money " or " the amount of unemployment," and in such cases we have to define and measure an allied attribute, such as the level of prices or the number unemployed according to some chosen definition of unemployment. (2) In choosing samples, the rule generally followed is to take only those where the definition is adequate and the measurement accurate, and in the best known index-numbers the choice is then so limited that all quotations which satisfy the rule are included. It very often happens that in this way the definition of the group must be reconsidered and limited. Thus if we start out to measure prices in general, the necessities of definition generally limit us to wholesale prices of goods which have regular market quotations ; and in wages the Labour Department is limited to cases where wages or rates are agreed on or standardised (except in the case of agriculture). In order that the resulting index-number should be subject to the analysis of the law of error the samples should be random and independent in their fluctuations from the general movement ; dependence increases the number of samples necessary for an assigned precision. Randomness

may, perhaps, be secured by the accidents which make the samples eligible; this is probably the case with wholesale prices but not with wages. Where the selection is biassed, we may sometimes obtain safety by further restriction of the definition of the group. (3) If the number of independent quantities is at all considerable, any reasonable system of weights is likely to give as good a result as the conditions of the problem allow.

Suppose that the changes in a group of quantities are determined by one general force which acts on all in the same sense, that is, tends to increase all or decrease all, and by several other forces each of which acts on one or more of the quantities, and some of which tend to increase, others to decrease the quantities they affect; then of the special forces, some wil' tend to increase, others to diminish the average, while the general force will have a cumulative effect entirely towards increasing, or entirely towards diminishing it. If the separate effects of the special forces are small compared with their number, they will tend to neutralise one another in their influence on the average; and the change in the average will show the influence of the general cause only. In the language of the last chapter, the special forces produce unbiassed changes, which are negligible in their effect on an average, in comparison with the biassed changes produced by the general force.

It appears from consideration of many of the index-numbers in ordinary use that the quantities actually measured are not those whose general movement we wish to know. Wholesale prices do not move with retail prices in accordance with any simple law, either of constant difference or constant ratio; standard wages differ in an unknown way from average wages; piece-rates have a varying and unknown relation to earnings. We do not get any such simple relations between the quantity that can be measured and the property that is really in question as $y = x$, or $y = kx$, or $y = a + bx$; but rather $y = f(x)$, where the form of the function is unknown. In order that the index-number may be intelligible, $y = a + bx$ must be a good approximation over the ordinary range of x—for extreme values of x terms of higher powers may become important and the resulting index untrustworthy. Here a disappears in the process of forming an index-number. It is often difficult to determine b, which measures the ratio of a change in y to that of a change in x.

The resulting index-number for any assigned year may be defined and expressed thus : Let $x_1, x_2 \ldots x_n$ be n quantities whose general movement is to be studied.   Let $y_1, y_2 \ldots y_r \ldots y_n$ be measured quantities related to the former by equations of a form to which $y_r - 100 = b_r(x_r - 100)$ is a good approximation.* Let suitable weights $w_1, w_2 \ldots$ be assigned and write J for $\dfrac{w_1 x_1 + w_2 x_2 + \cdots}{w_1 + w_2 + \ldots}$, and I for $\dfrac{w_1 y_1 + w_2 y_2 + \cdots}{w_1 + w_2 + \ldots}$. Then J is the incalculable theoretic index-number whose changes express the movement, and I is the index-number calculated and used.

$$I - 100 = \frac{\Sigma w (y - 100)}{\Sigma w} = \frac{\Sigma wb (x - 100)}{\Sigma w}.$$

Let $b_1 = k + d_1, b_2 = k + d_2 \ldots$, where k is chosen, if possible, as that average of the b's which makes

$$\frac{\Sigma wd (x - 100)}{\Sigma w} (= F, \text{say})$$

small for the ordinary range of values of the x's.
    Then

$$I - 100 = k \frac{\Sigma w (x - 100)}{\Sigma w} + \frac{\Sigma wd (x - 100)}{\Sigma w} = k (J - 100) + F.$$

If the x's have, in general only, a moderate range of values, if the b's are nearly equal and extreme values of b do not coincide with extreme values of w, then F is small and its variation from year to year negligible.

In this case I is so related to J that it equals 100 in the standard year when J (and every x and y) is 100, and a change in its value is very nearly k times the change in J, where k is an average of the b's which measure the ratio of the changes of the various y's to those of the corresponding x's.

If we try to make a retail price index-number out of wholesale prices, the b's are not known, and presumably differ greatly from one commodity to another, from time to time, and vary in an unknown way when prices are specially high or low.   Hence the connection between general retail prices and wholesale prices is not so close as to allow the statement that a change in the one is directly proportional to a change in the other.   In the case of the Labour Department's index-number of wages, the changes in time-rates have not the same relation to earnings as have those in piece-rates, and in neither group is the relation

---

* This equation can readily be obtained by a rearrangement from $y = a + bx$.

known; that is, the b's are unknown and are not equal. So far as piece-rates are concerned, when these rates are rising it often happens that earnings rise more rapidly (b less than 1), and when they fall that the earnings fall less rapidly (b greater than 1); that is, the b's are not constant and F, in the formula just given, is unknown and not negligible.

If the b's are equal, F is zero, and k could be determined by special examinations in two years. Then the movements of I would reflect faithfully the movements of J on a known scale.

The actual relations are not known; if an x is 4 per cent. above the average, we do not assume that y is also 4 per cent. above its average, but assume that its deviation is 4 per cent. $\times b$, where b is nearly constant. The b's differ from some (weighted) mean value (k), and it is assumed that the effect of these differences nearly disappears when the average is taken, and that the mean value, k, is nearly constant from year to year. Various hypotheses can be made as to the values of the b's and the resulting value of k, and the fluctuations of the index-numbers interpreted.

It is essential that when an x returns to a value after a fluctuation, the corresponding y shall return to its former value, or at least that any differences shall be small and unbiassed. This condition would be broken if wholesale prices were used to measure the changes in retail prices, while the relation between the two gradually changed, as presumably it does. It is broken in the Labour Department's index of the general course of wages, in so far as changes in standard wages or piece-rates have a varying relation to changes in average wages.

There are many index-numbers of wholesale prices, extant, some of which we may pass in review. The Board of Trade publish the recorded quantity and value of goods imported and exported, and the average prices of these goods can be calculated. Those commodities are selected which occur in the returns for the whole period chosen. A particular year is chosen as base; then the goods are valued in all other years separately at their prices in the base year; the total of these values in any year is the sum which the goods would have been worth if their prices had remained unchanged; the ratio of this value to that actually recorded is the ratio of their average price in the base year to their average price in the other year selected (if the term average is used broadly), and if the first term of this ratio is equated

to 100, the second term is the index-number required for the year selected, expressed as a percentage of the number for the base year.   It is at once evident that we are here dealing with weighted averages.

Let p_1, p_2, p_3 . . . be the prices in the base year of units of the goods selected, and r_1p_1, r_2p_2, r_3p_3 . . . the prices in the year for which we require an index-number : then r_1, r_2, r_3 . . . measure the changes of prices for the separate commodities, and these r's are the samples from which we are to deduce the general change of price. The weights used in the process described may, be found thus : let b_1, b_2, b_3 . . . be the numbers of units of goods in the selected year; then the total value in the selected year at the prices of that year is $(b_1r_1p_1 + b_2r_2p_2 + . . .)$, and at the prices of the base year is $(b_1p_1 + b_2p_2 + . . .)$; the ratio is $\Sigma brp : \Sigma bp$, and the index-number for the selected year is

Systems of weights.

$$100 \times \frac{\Sigma brp}{\Sigma bp} = 100 \times \Sigma\left(r \cdot \frac{bp}{\Sigma bp}\right).$$

Here the weights applied to the r's are the values which the corresponding goods in the selected year would have borne at the prices of the base year.   It is clear that the selection of the standard year affects the weights, for any particular commodity can be given special weight by choosing as base a year in which its price is high, and much trouble has been spent in searching for a " normal " year; but though the weights of separate commodities are affected, it does not follow that the average will be altered, and we should expect from the principle laid down above that the change would be very slight.   In fact we have the following figures :—

INDEX-NUMBERS OF 1886 AND 1883 COMPARED.*								
	IMPORTS.				EXPORTS.			
Weights. {	Values at 1873 Prices.	Values at 1883 Prices.	Values at 1861 Prices.	Values at 1881 Prices.	Values at 1873 Prices.	Values at 1883 Prices.	Values at 1861 Prices.	Values at 1881 Prices.
1883	100	100	100	100	100	100	100	100
1886	81.7	82.1	82.9	82.3	88½	88	87	89

   * From the *Economic Journal* and the *Statistical Journal*, both June 1897.

It is possible to produce figures which show a variation caused by a change of base year, but it is done by choosing samples which lend themselves to the special argument.

• . Since so great an alteration in choice of weights makes so little difference, it is worth while to see if we need even keep the weight due to the quantities imported (the b's in the above formulæ). The following table may be quoted * to show that these weights even have little influence :—

Index-Numbers for 1895, *when that of* 1881 *is* 100, *obtained by Various Systems of Weighting.*

	RATIOS OF PRICES $(r_1, r_2 \ldots)$					Reciprocal of A.M. of $\frac{1}{r_1}, \frac{1}{r_2}$ &c.	*Economist's* Figures.
	Weighted by Values of 1895 Quantities at 1881 Prices.	Weighted by Declared Values in 1881.	Arithmetic Mean.	Median.	Geometric Mean.		
Imports	$67\frac{1}{2}$	69	$73\frac{1}{2}$	$72\frac{1}{2}$	$72\frac{1}{2}$	69	71
Exports	83	87	82	81	$78\frac{1}{2}$	75	

Let $b_1, b_2 \ldots$ be quantities and $p_1, p_2 \ldots$ prices in 1881, and let $c_1, c_2 \ldots$ be quantities and $r_1 p_1, r_2 p_2 \ldots$ prices in 1895.

The first column gives the result of—

$$100 \times \frac{\text{Sum of 1895 quantities at 1895 prices}}{\text{Sum of 1895 quantities at 1881 prices}}$$

$= \dfrac{100 \, Sc_1 p_1 r_1}{Sc_1 p_1}$, and the weights applied to the r's are the 1895 quantities valued at the 1881 prices.

The second column gives the result of—

$$100 \times \frac{\text{Sum of 1881 quantities at 1895 prices}}{\text{Sum of 1881 quantities at 1881 prices}}$$

$= \dfrac{100 \, Sb_1 p_1 r_1}{Sb_1 p_1}$, and the weights applied to the r's are the declared values of 1881.

In the next three columns the arithmetic mean, the median, and the geometric mean of the r's are given. In the last column but one the arithmetic mean of $\dfrac{1}{r_1}, \dfrac{1}{r_2} \ldots$, that is of the ratios

---

* From the *Economic Journal* (with a correction in the statement of weights).

of the prices of 1881 to 1895, is calculated, and the ratio of this mean to 100 equals the ratio of 100 to a new index-number, which corresponds to the former arithmetic mean with the years 1881 and 1895 interchanged. The figure in the last column is calculated from material given in the *Economist ;* every year the imports and exports are valued at their prices in the previous year, and thus an annual ratio is given similar to that in the first column of figures in the table just given; the number 100, taken for 1881, is multiplied by this annual ratio year by year till 1895, and the number 71 is the result. [Algebraically this index-number is :—

$$100 \times \Sigma\left(r \cdot \frac{bp}{\Sigma bp}\right) \times \Sigma\left(r^1 \cdot \frac{b^1 p^1}{\Sigma b^1 p^1}\right) \times \times .]$$

A more complete analysis of these figures, and an investigation as to the causes of the divergence between the export indices 87 and 75, would show which of the methods should be adopted. Here we will be content with noticing that the unweighted average, 82, is very near the first weighted average, 83.

Further methods of dealing with such weights are given on pp. 209–211, under Retail Index-Numbers.

The advantage of index-numbers on the Board of Trade basis is that they measure approximately an objective quantity, Objective measure. and a result is obtained which can be stated in terms which appeal to the ordinary man who is not a statistician : such as, " The imports of 1895 would have cost half as much again if their prices had been those of 1881 ; " but it does not follow that this index is the best measure of the less-definable quantity, " Fall in the price of imports," where we imagine a general cause affecting this class of commodities whose action is modified by other partial causes.

It is important to choose a normal year or the average of a Choice of base year. period as base, for the choice of year affects the effective weights in subsequent comparisons. Using the following notation—

Weights chosen.	Price in Base Year.	Price in Second Year.	Price in Third Year.	Ratio of Prices in Third and Second Years.
w_1	100	$100r_1$	$100r_1^1$	$R_1 = r_1^1 : r_1$
w_2	100	$100r_2$	$100r_2^1$	$R_2 = r_2^1 : r_2$
⋮	⋮	⋮	⋮	⋮

and writing 100, I_1, I_2 as the index-numbers in the three years, we have—

$$\frac{Sw100}{Sw} : \frac{Sw100r}{Sw} : \frac{Sw100r^1}{Sw} = 100 : I_1 : I_2$$

$$\therefore I_2 = I_1 \times \frac{Swr^1}{Swr} = I_1 \times \frac{S(wr \cdot R)}{Swr}$$

Whereas if we had taken the prices all as 100 in the second year we should have $I_2 = I_1 \times \dfrac{SwR}{Sw}$.

If the averages were unweighted we should still have the same difficulty, for then the values would be $\dfrac{SrR}{Sr}$ on one system and $\dfrac{1}{n}SR$ on the other.

Since errors in weights have under ordinary circumstances but little effect, it is only when a quite abnormal base year is chosen, or when prices are moving very irregularly, that this consideration becomes important.

Professor Edgeworth has pointed out that the use of the geometric mean avoids this difficulty in the case of unweighted averages. In the same notation—   *Geometric mean.*

$$100 : I_1 : I_2 = 100 : 100\sqrt[n]{r_1 r_2 \ldots r_n} : 100\sqrt[n]{r_1^1 r_2^1 \ldots r_n^1}$$

$$\therefore I_2 = I_1 \times \sqrt[n]{\frac{r_1^1 r_2^1 \ldots r_n^1}{r_1 r_2 \ldots r_n}} = I_1 \times \sqrt[n]{R_1 R_2 \ldots R_n,}$$

so that the same result is obtained for the comparison of two years whatever year is taken as base.*

Mr. Sauerbeck and the *Economist* both avoid in part the difficulty of weighting the separate ratios by their relative importance in consumption, by selecting from those   *Other index-numbers.* commodities whose prices are most accurately determined more instances of such widely consumed articles as wheat than of less important commodities such as linseed. Mr. Sauerbeck has, in his annual articles in the *Journal of the Royal Statistical Society*, verified the correspondence of the unweighted average of his 45 ratios with the average of the same weighted on various principles.†

While the choice of the special weights to be employed is,

---

* On this point, and on others in this chapter, see article Index-Numbers, in Palgrave's *Dictionary of Political Economy*.

† See, for example, *Statistical Journal*, 1900, pp. 97, 98.

when the number of ratios taken is at all considerable, quite
Importance of right choice of samples. unimportant, the choice of the quantities dealt with has great effect on the result. Thus import figures, relating to raw materials and the produce of other countries, do not lead to the same index-numbers as export figures dealing with the price of our own produce, though the tables just given show that they are little affected by weights ; and neither of these agree closely with Mr. Sauerbeck's or the *Economist's* numbers, and these again are not in complete agreement. The samples on which these four sets of numbers are based are from different groups of commodities, and the numbers show that the same forces do not affect these groups in the same degree. When we have so multiplied our samples, that we can subdivide them without affecting the index-numbers deduced, we may expect our results to represent the required measurement.*

If we compare the *Economist* index-numbers with Sauerbeck's during the period 1860–70, we see that the former show
Great advantage of the median. a very much greater increase during the cotton famine than the latter. An index-number which can be greatly disturbed by fluctuations, however violent, in only one group of commodities, is clearly wanting in some of the chief qualities of a general measure of price levels. A very simple means of avoiding this difficulty, and indeed all the intricacies of weighting, is to take the median of all the price ratios of a particular year as the index-number of that year. It is perhaps impossible to show theoretically that any other average satisfies the required conditions better than the median, if a sufficient number of items are included, and there can be no doubt that it is practically the easiest to calculate.

If, on the other hand, paucity of data makes the inclusion of weights necessary, and the popular desire for concrete measure-
Proposed standard. ments makes a fine show of weighting expedient, we perhaps cannot do better than to adopt such a standard as that proposed by the Committee of the British Association, for the construction of an index-number, which might be the basis of business transactions involving future payments. This standard is as follows :—

---

* Mr. Sauerbeck's numbers are to be found in annual articles by him in the *Statistical Journal* ; and a diagram showing them from 1820 is published by P. S. King & Son.

Basis of Index-Number recommended by the Committee appointed by the Economic Section of the British Association, 1888.

Articles.	Estimated Expenditure per Annum on each. 000,000's omitted.	Hence Weights assigned.	Prices to be taken from
Wheat · · ·	£60	5 ⎫	Gazette average, English wheat.
Barley · · ·	30	5 ⎬ 20	" " barley.
Oats · · ·	50	5	" " oats.
Potatoes, rice, &c. ·	50	5 ⎭	Av. import price, potatoes.
Meat · · ·	100	10 ⎫	Market quotations, live meat, Smithfield.
Fish · · ·	20	2½ ⎬ 20	Board of Trade Returns; average per cwt. landed.
Cheese, butter, milk	60	7½ ⎭	Cheese and butter, average import price.
Sugar · · ·	30	2½ ⎫	Av. import price, refined sugar.
Tea · · ·	20	2½	" " " tea.
Beer · · ·	100	9 ⎬ 20	" export " beer.
Spirits · · ·	40	2½	" import " spirits.
Wine · · ·	10	1	" " " wine.
Tobacco · · ·	10	2½ ⎭	" " " tobacco.
Cotton · · ·	20	2½ ⎫	" " " cotton.
Wool · · ·	30	2½	" " " wool.
Silk · · ·	20	2½ ⎬ 10	" " " raw silk.
Leather · · ·	10	2½ ⎭	" " " hides.
Coal · · ·	100	10 ⎫	" export " coal.
Iron · · ·	50	5 ⎬ 20	Market price, Scotch pig-iron.
Copper · · ·	25	2½	Av. import price, copper ore.
Lead, zinc, tin ·	25	2½ ⎭	" " lead ore.
Timber · ·	30	3 ⎫	Average import price.
Petroleum · ·	5	1	" " "
Indigo · ·	5	1 ⎬ 10	" " "
Flax and linseed ·	10	3	" " "
Palm oil · ·	5	1	" " "
Caoutchoux · ·	5	1 ⎭	" " "

American statisticians have adopted a method of comparing totals instead of weighted or unweighted price-ratios for the formation of index-numbers. " By so doing, it is maintained, two difficulties are overcome : First, the problem of choosing a base year, since actual prices do not necessarily have to be reduced to a relative basis, and, second, of deciding on an appropriate average of relatives." [*] In fact the method, though it may have advantages in intelligibility and simplicity of construction, introduces no new principle. It may be thus described :—The price of each article in, say, 1914 is multiplied

[*] Secrist, *An Introduction to Statistical Methods*, 1917, pp. 329 and 339, 340. See the *Bulletin of the United States Bureau of Labor Statistics*, Whole Number 181, October 1915.

by the quantity marketed in the last census year, 1909; the price in, say, 1912 is multiplied by the same quantity. With the aggregate for 1914 as the base, or 100, the index-number for 1912 is obtained by comparing the 1912 aggregate with the 1914 aggregate. If $w_1, w_2 \ldots$ are the quantities, $P_1, P_2 \ldots$ the prices in 1914 and $p_1, p_2 \ldots$ those in 1912, the aggregates are SwP, Swp, and the index is $100 \dfrac{Swp}{SwP} = 100 \dfrac{S(wP \cdot R)}{SwP}$ where $R_1, R_2 \ldots$ are the price ratios 1914 to 1912. This is equivalent to the Board of Trade index discussed above, and has no special claim to accuracy.

Since we can only obtain rough correspondence in dealing with wholesale prices, we cannot expect to be able to measure **Retail price index.** retail prices with any great precision. For we saw in the preceding chapter that the error in an average bears a definite relation to the errors in the items which compose it; if the errors in the items are on the whole doubled, it is likely that the errors in the average and in the ratio of two averages will also be doubled, and we shall need four times * as many samples to restore the precision. Unfortunately the material for computing a retail index-number is even more incomplete than that for wholesale prices, and owing to the smaller number of articles that can be included, and the preponderance of such items as bread and rent, the question of weighting becomes of more importance.

When we wish to construct an index-number to show the purchasing power of money of special classes, we must take **Special difficulties.** into account some considerations which can be ignored when dealing with wholesale price numbers. Different classes of persons at the same time, and the same classes at different times, spend their income in varying proportions on different objects. If we could collect enough sufficiently accurate samples, this fact would not matter so much; but it would still be of some importance owing to the tendency to make increased purchases of cheapening commodities. As it is, it would be necessary to construct separate index-numbers for each class and each district. The difficulty of insufficient and inaccurate data cannot at present be overcome; but as it is possible that we may in the future get definite records of retail prices sufficiently numerous to make up for

---

* See Part II, Chap. IV.

their want of precision, we may glance at the other details of the problem. To form an index-number for a particular class of people, we need records of the method of expenditure of their income at all the dates in question, of sufficient numbers to obtain the slight precision which weighting needs. Then if we had fairly good records of retail prices several methods of weighting are open to us,[*] all of which are likely to give nearly the same result. The necessity of weighting and the methods are best shown by a numerical illustration.[†]

Methods of weighting.

The data for the measurement of the change of the cost of living, however it is defined, are always of the same nature, and consist of records of the quantities of various commodities bought and the prices paid for them at two dates or places or by representatives of different social groups. Thus we have given with greater or less accuracy :—

Commodity.	Place or Date. A.			B.		
	Quantity.	Price.	Expenditure.	Quantity.	Price.	Expenditure.
1	Q_1 ×	P_1	= E_1	q_1 ×	p_1	= e_1
2	Q_2 ×	P_2	= E_2	q_2 ×	p_2	= e_2
3	Q_3 ×	P_3	= E_3	q_3 ×	p_3	= e_3
⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	Q_n ×	P_n	= E_n	q_n ×	p_n	= e_n

In the Table on p. 210 are shown in this form the budgets used in the Report of the Committee on Cost of Living, 1919.

The second year's budget at the first year's prices would have cost $225.5d$. instead of $455.5d$. The index-number of retail prices on this basis is $100 \times \dfrac{455.5}{225.5}$ or $100 \dfrac{Sqp}{SqP} = 202.0$. The weight applied to a price ratio $p : P$ is qP. The index-number $= 100 \dfrac{Se}{Se\frac{P}{p}}$ . . . . (a)

The first year's budget at the second year's prices would cost $521.6d$. instead of $246.5d$. The index-number on this basis is

---

* See article on Wages, Nominal and Real, in Palgrave's *Dictionary of Political Economy*, pp. 640–41.
† Taken with part of the context from " The Measurement of Changes in the Cost of Living," *Statistical Journal*, 1919, pp. 343 *seq*.

$$100 \times \frac{521 \cdot 6}{246 \cdot 5} \text{ or } 100 \frac{SQ p}{SQP} = 211 \cdot 6 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (b)$$

The weight applied to a ratio $p : P$ is QP or E. This is the method used in the *Labour Gazette* to measure the " average increase in retail prices."

Urban Working-class Budgets. (Based on Cd. 8980, p. 18.)

EXPENDITURE OF STANDARD FAMILY.

			1914.			June, 1918.			p/P
			Q Quantity.	P Price.	E Expenditure.	q Quantity.	p Price.	e Expenditure.	Price Ratio.
				d.	*d.*		*d.*	*d.*	
1.	Bread and flour	lbs.	33.5	1.51	50.5	34.5	2.36	81.5	1.56
2.	Meat - -	,,	6.8	8.6	58.5	4.4	18.6	82.0	2.15
3.	Bacon - -	,,	1.2	11.7	14.0	2.55	26.1	66.5	2.24
4.	Lard, suet, etc. -	,,	1.0	7.5	7.5	.78	17.9	14.0	2.29
5.	Eggs - -	No.	13	1.0	13.0	9.1	4.0	36.5	4.00
6.	New milk -	pints	9.2	1.8	16.5	11.7	3.0	35.5	1.69
7.	Condensed milk	tins	.25	6.0	1.5	.59	14.5	8.5	1.42
8.	Cheese - -	lbs.	.84	8.9	7.5	.41	20.7	8.5	2.32
9.	Butter - -	,,	1.70	14.4	24.5	.79	29.7	23.5	2.07
10.	Margarine -	,,	.42	6.0	2.5	.91	12.1	11.0	2.01
11.	Potatoes - -	,,	15.6	.7	11.0	20	1.25	25.0	1.78
12.	Rice and tapioca	,,	1.4	3.2	4.5	1.3	5.8	7.5	1.82
13.	Oatmeal - -	,,	1.3	.1.9	2.5	1.4	4.3	6.0	2.24
14.	Tea - -	,,	.68	21.3	14.5	.57	33.3	19.0	1.56
15.	Coffee - -	,,	.09	16.7	1.5	.12	25.0	3.0	1.50
16.	Cocoa - -	,,	.18	19.4	3.5	.23	32.6	7.5	1.69
17.	Sugar - -	,,	5.9	2.2	13.0	2.83	7.07	20.0	3.21
	Total - - -		—	—	246.5	—	—	455.5	—
	Other food - -		—	—	52.5	—	—	111.5	—
	Total - - -		—	—	299.0	—	—	567.0	—

S. QP=246.5   S. qp=455.5   3. Qp=521.6   S. qP=225.5
S. $e \div$S. E=1.90   S. $Qp \div$S. QP=2.12   S. $qp \div$S. qP=2.02

In some cases there may be reasons for preferring (*a*) or preferring (*b*). If not, it is reasonable to take a mean between the results; the arithmetic mean is 206·8, the geometric mean is 206·74, the harmonic mean 206·69, and it is usually indifferent which we take. Or a method which may be commended for its simplicity in idea is to take the averages of the quantities seriatim ($\frac{1}{2} \overline{Q_1 + q_1}, \frac{1}{2} \overline{Q_2 + q_2} \ldots$) and find their cost in each year and compare their sums. This gives $\dfrac{S(Q + q)p}{S(Q + q)P} \times 100$ = 203·7. The weight applied to a ratio is now $\frac{1}{2}(Q + q)P$ (*c*)

Another method is to take the average of the expenditures at the two dates on each item as the weight for the price ratio of that item, so obtaining $\dfrac{S(E + e)\frac{p}{P}}{S(E + e)} = 198\cdot6$. ． ． ． (d)

This, however, involves the quantity p^2/P in the numerator and gives undue weight to exceptional movements of prices of particular commodities.

In absence of knowledge of quantities the simple average of the price ratios $\dfrac{1}{n}S\dfrac{p}{P} \times 100 = 209\cdot1$ ． ． ． ． ． ． (e)

is sometimes taken; but it is never safe to neglect weights in this problem, though it is not necessary to aim at great precision in them.

Finally, a more complicated method has been advocated in which it is supposed that the second total is expended in the same proportion item by item as the first, and the quantities of each item thus purchasable are valued at the price in the first year. The ratio of the whole actual expenditure in the second year ($\times$ 100) to the expenditure so calculated

$$= \frac{100\,Se}{S\left\{E_1\dfrac{Se}{SE} \times \dfrac{P_1}{p_1} + \ldots\right\}} = 100\,\frac{SE}{SE\frac{p}{P}} = 196\cdot4 \quad \text{．．．} \quad (f)$$

The weight applied here to the ratio $\dfrac{p}{P}$ is $QP^2 \div p$, and as in case (d) gives undue weight to particular prices. Also there is no reason to suppose that the expenditure is kept in a constant ratio item by item.

No agreement has been reached on the question which method is the best for the measurement of retail prices; but there are·serious theoretical objections to (d) (e) (f). There is nothing in general to choose between (a) and (b), but for this purpose one year has the same claim to be included as the other and we are therefore obliged to take a mean. Of the various means the method (c) of averaging the quantities is the most sensitive, is quite easy to compute, and on all grounds is to be·recommended.*

The problem of measuring the movement of retail prices has been generally confused with that of measuring the change

---

* This opinion is different from that expressed in former editions. For further information see the bibliography in the article on Workmen's Budgets in Palgrave's *Dictionary of Political Economy.*

in cost of a standard (representing either minimum subsistence or efficiency subsistence) with the items the same at both dates. It is not proposed here to discuss such a measurement in detail, but it should be realised that there is a continual change in the prices and supply of the various commodities. For such budgets it ought to be assumed that the same nourishment (or more generally the same satisfaction) is obtained at each date by the most economical purchases, so that the quantities of those foods whose price has risen least or fallen most are increased while others are diminished, and consequently an upward movement is less and a downward movement greater than that measured by method (a).*

There are still two further considerations which hinder the complete solution of the problem. In all budgets rent is

Further difficulties. an important item, and there seems no prospect of obtaining any good estimate of the relation between increasing rent and improving accommodation, allowing for the benefits of public expenditure paid by rates included in rent. Again, if we consider, not how money is spent, but how it might be spent, we should have to introduce a more general factor; for the margin which remains when necessities are satisfied may have a rapidly growing purchasing power, as the products of machinery increase in variety and diminish in price; perhaps the calculated fall in wholesale prices forms a fair measure of this growth.

Leaving this very difficult problem, let us return for a moment to the measurement of a quantity more typical of index-numbers.† If we have to measure the action of a cause,

Index-numbers of consumption. which affects quantities which have no common measure, we are still able to apply index-numbers. A general increase has taken place in the consumption of imported goods, and if we can measure this increase independently of any change in price, we can use it for criticism of any measurement of a movement in real wages. The only common measure of bread, currants, cheese, meat, etc., of practical value is their price, their weight being useless for the purpose; consequently another method is necessary. If the quantities

---

* For the discussion of these questions see "Cost of Living," *Statistical Journal*, May 1919.
† The following illustration is based on Mr. G. H. Wood's paper on "Some Statistics of Working Class Progress," *Statistical Journal*, 1899.

consumed year by year of a number of such commodities are written down, expressed as percentages of the consumption in any years (not necessarily the same), we have series of numbers which only need weighting to form the index-number required. We can in this case verify, that any logical choice of weights, based on their value or their assumed importance, or even a random system of weights, gives much the same index-number as the simple arithmetic averages; in fact, we have a sufficiently good group of samples to render us nearly independent of weights. When this is the case we can say with safety that the number required lies in the neighbourhood of the group given by the various systems of weights, and choose what appears the most logical system for the estimate we adopt. In the paper referred to, five different systems applied to only fourteen commodities give results for the increase of consumption all between 13·8 and 20·1 per cent. in the period 1873-96.

The application of index-numbers to wage statistics does not involve any fresh principles. It is not permissible to ignore the change of weights in this case; for otherwise   Wage index-
we should not allow for the general tendency to   numbers.
increase numbers where wages are rising. There is great liability to "biassed" errors in separate averages; for wages for overtime, specially high piece-wages, wages of large uncombined classes of low-skilled or badly paid workpeople, may often be omitted in wage records. These biassed errors, however, tend to disappear in comparison; and it may prove possible to construct a wage index-number of very fair precision.*

. *Note added in* 1936.—Write I_1 S(QP) = S(Qp), I_2S(qP) = S(qp), in the notation of p. 210, and J_1S(QP) = S(qP) for a corresponding index of quantity.
Then

$$SQP\left(\frac{q}{Q} - J_1\right)\left(\frac{p}{P} - I_1\right) = S(qp) - J_1S(Qp) - I_1S(qP) + I_1J_1S(QP)$$

$$= I_2S(qP) - I_1J_1S(QP) - I_1S(qP) + I_1J_1S(QP)$$

$$= (I_2 - I_1)S(qP)$$

∴ when an *increase* of price in a commodity greater than that measured by I_1 goes with a *fall* in quantity consumed, as compared with J_1, I_2 is less than I_1. Part of the rise of prices may be expected to be evaded in this way, *e.g.*, on p. 210, I_2 = 2·02, I_1 = 2·12. [Cf. *International Comparisons of Cost of Living*, I.L.O., Geneva, 1934, pp. 15-16.]

* For a complete illustration of method and of the various factors involved, see "The Statistics of Wages in the United Kingdom. Part XIV.: Engineering and Shipbuilding," *Statistical Journal*, March 1906, pp. 154 *seq.*, especially pp. 166, 168 and 185.

CHAPTER X.

INTERPOLATION.

SECTION I.—GENERAL.

IT is very often the case in practical statistics that we are
not able to make serial estimates as frequent or descriptions
Necessity of of groups as detailed, as is necessary for their use
Interpolation. in further investigations. Thus the population is
only counted once in ten years; but we need to bring monthly
and annual accounts—births, deaths, trade returns, etc.—into
close relation to the existing number of people, and estimates
for the budget and the yield of taxes must be based on the
assumed number of taxpayers for the current year; it is
therefore necessary to *interpolate* estimates for the number of
the people in intercensal years. Again, interpolation is needed
for the statement of the distribution of the population accord-
ing to age, a tabulation which is necessary for actuarial work
and for sociological purposes. The ages returned on the
householder's schedule are nominally correct to the year, but
in practice they are known to be inaccurate, tending to group
themselves in the neighbourhood of round numbers; but the
returns for such age periods at 35–45 years are more correct,
since the persons who return themselves as 40 years old are
probably within 5 years of that age. The original returns
are so erroneous that prior to 1911 they were not published,
but the numbers were only given in the ten-yearly periods;
from the numbers so given, it is necessary to estimate the
numbers for the individual years. Again, the compilers of the
wage census of 1886–91 enumerate the numbers earning wages
" of 15s. and under 20s.," " of 20s. and under 25s.," and so
on, but not the numbers in shilling limits. In problems
relating to wages we often need more detail; and when we are
comparing these wages with a similar group in France, we
must devise a scheme by which grades of 2 francs can be com-
pared with grades of 5s., by a suitable system of interpolation.

Such a necessity is very common when we wish to compare groups, which are similar but tabulated on diverse systems. Thus, two countries conduct their census at different dates. In one country the age groups are of fifteen years, in another of ten; in one, "young persons" are those under 21; in another, those under 18. Occasional estimates seldom correspond in date; wage statistics are found for 1840, 1850, and 1892 in France, and for 1866, 1885, 1886, 1891, and 1906 in England. Similar differences are found when we are comparing county with county; and a discussion of the method of determining averages in such a case will illustrate some of the elementary problems of interpolation.

Suppose that the figures printed in Roman type in the following table are accurate returns of the weekly wages in three districts, and that we wish to find the average change in the three together. *Elementary example.*

Years.	1860.		1862.		1864.		1866.		1870.		1871.		1875.		1878.		1880.		1881.	
	s.	*d.*	*s.*	*d.*	*s.*	*d.*	*s.*	*d.*	*s.*	*d.*	*s.*	*d.*	*s.*	*d.*	*s.*	*d.*	*s.*	*d.*	*s.*	*d.*
District A	12	6	15	0	*15*	*0*	15	0	*15*	*0*	14	6	18	0	18	0	17	6	17	0
„ B	18	0	*19*	*0*	19	0	*20*	*0*	20	0	*19*	*6*	*21*	*0*	21	0	*20*	*6*	20	0
„ C	10	0	*11*	*0*	11	0	12	0	*12*	*0*	*12*	*0*	15	0	*15*	*0*	15	0	*14*	*6*
Average	*13*	*6*	*15*	*0*	*15*	*0*	*15*	*8*	*15*	*8*	*15*	*4*	*18*	*0*	*18*	*0*	*17*	*8*	*17*	*2*

It is clear that there is something to be learnt about the general course of wages from the data, but the lessons are not obvious.. The following figures, printed in the table in italics, are those which naturally suggest themselves. There is no sign in A of any change between 1862 and 1866, so we write *15s.* for 1864. Judging from B, the figure for 1870 is not likely to have been lower than that for 1864, so we write *15s.* for A in 1870. A is now complete; we notice that in A the first rise was complete by 1862, and assuming the same in B, we obtain *19s.* for 1862. In C there is a rise between 1864 and 1866, while in A there is no change from 1866 to 1870; B will correspond if we write *20s.* in 1866. If we write for B, *19s. 6d.* in 1871, *21s.* in 1875, and *20s. 6d.* in 1880, we shall have close correspondence with A from 1866 to 1881. Similar

reasons lead to the numbers interpolated for C. The un-
weighted average can then be calculated year by year, which
could not be done directly from the date. This average
reflects all the changes in the original figures and gives no special
predominance to any. It may be regarded as the most probable
series that can be based on the given information.

We will now notice the assumptions tacitly made in pro-
ceeding by this method. First, it has been assumed that
there are no sudden jumps, that such a figure as
20s. for A 1864 is inadmissible ; this is only justifi-
able if we are acquainted with the general causes which influence
the rate of wages, and know that there was no violent disturb-
ance in the intermediate dates. We could not make this
assumption as to wages in the cotton trade in the time of the
American Civil Wars, nor can we make it over a long series
of years. Secondly, it has been assumed that in the absence
of evidence to the contrary the rise or fall has been uniform.
Thus, in B 1878–81, the wage in 1880 is assumed to be inter-
mediate between 1878 and 1881 ; if there had been no indica-
tion from A that it was half-way between in point of wages,
it might have been said that in point of time it was two-thirds
of the way, and 20s. 8d. should be interpolated for 1879 and
20s. 4d. for 1880, if it was worth while to depart from round
numbers. Thirdly, it has been assumed that the course of
wages in the three districts was similar. Thus in A there is
a rise from 1860–62, but there is no further improvement at
any rate before 1866 ; it is consequently assumed that the rise
registered in B and C before 1864 actually took place before
1862. Again, when considering the period 1870–75, we notice
that in A there is a fall till 1871, and a sharp rise to 1875, and
no change to 1878 ; in B, therefore, it is assumed that the wage
of 1875 is equal to that of 1878, and the fall in 1878 may be
allowed because it increases the sharpness of the rise in 1871–75.
In C it is doubtful whether the 12s. in 1871 should not rather
be 11s. 6d. The reasons against are that a gain on a low wage
is often not so easily lost as a gain on a high one; 6d. is a
larger drop proportionately on 12s. than on 15s.; that the rise
of 3s. 6d. which would then be shown 1871–75 is a larger
proportionate rise than in either A or B; and that the exist-
ence of the fall in 1870–71 depends only on the evidence of a
fall between 1866–71. When the figures are few in number,

Assumptions made.

it is necessary to examine them in this way to pick out the most probable; and it is often fairly easy to fill in tHe figures which satisfy all the existing evidence fairly closely.

The question at once arises, What certainty have we that these quantities, by hypothesis unknown, are in reality anywhere near the figures which on the face are most probable?

In some cases of interpolation, dealt with presently, the answer can be given as a statement of mathematical probability, such as : it is 2 to 1 against a divergence of 6d. from the assigned figure, 30 to 1 against one of 1s., 1000 to 1 against one of 2s. 6d., and so on; but in the figures most often cropping up in investigations it is not possible to assign such a precise probability. There is one rough but useful way of testing the accuracy of such interpolation as in the case before us which can be explained by an example. Test how far we can throw out our calculated average for 1870, without violently infringing the common-sense of the question. Make A and C as large as possible in these dates; we may perhaps suppose a rise of 1s. above 1866, seeing that there is one in B between 1864 and 1870. We can hardly suppose either that 1870 is as high as 1875–78, or that there is a great drop of as much as 2s. in the single year, if we are acquainted with the causes that determine the wages at those dates. Let the highest wage we can assign to A and C be 16s. 6d. and 13s. 6d. respectively. Our average is then 16s. 8d. instead of 15s. 8d. Similarly, we might perhaps think that 14s. and 11s. were the lowest possible in A and C in 1870; then the average would be 15s. Assuming that we know enough about the general trend of events at these dates to assign limits in this way, we can say it appears improbable that the average wage in 1870 was less than 15s. or more than 16s. 8d., and that the evidence points to 15s. 8d.

The accuracy of our interpolation then depends—(1) On knowledge of the possible fluctuations of the figures, to be obtained by a general inspection of the fluctuations at dates fór which they are given; (2) on knowledge of the course of the events with which the figures are connected.

A second example of a similar kind * may be given to illustrate the numerical calculation.

Tests.

Numerical example.

* Taken from "Agricultural Wages in England," in the *Statistical Journal* December 1898, by the present author.

Northern Counties.					Weekly Agricultural Wages in			
					1867-69.		1869-70.	
					s.	*d.*	*s.*	*d.*
Cheshire	•	•	•	•	*13*	*1*	13	6
Lancashire	•	•	•	•	15	0	15	0
West Riding of Yorkshire		•	•	14	6	16	5
East	,,	,,	•	•	14	6	*14*	*11*
North	,,	,,	•	•	14	6	15	4
Durham	•	•	•	•	16	6	16	0
Northumberland	•	•	•	•	16	6	16	7
Cumberland	•	•	•	•	*14*	*4*	14	9
Westmoreland	•	•	•	•	*15*	*7*	16	1

Roman figures given.   Italic figures interpolated.

The averages of the wages in the five districts for which data exist in both periods are 15*s*. 4·8*d*. in 1867–69 and 15*s*. 10·4*d*. in 1869–70, that is in the ratio 33 : 34. If we assume that the wages in the other counties have been influenced by similar causes and increased in the same ratio, we obtain the figures interpolated in the table. The unweighted averages for the northern counties are now 14*s*. 11*d*. and 15*s*. 5*d*. in the two periods, instead of 15*s*. 3*d*. and 15*s*. 5*d*., the averages of the given numbers. For general comparison all over England between these two years we should have been obliged to neglect the missing counties in both years, which would have unfairly lowered the general average, since these counties have in recent times had wages above the English average though below that of the northern district. At the same time we should have unfairly raised the apparent average of the northern district. We should also have lost the probable figures for the special counties at the earlier date which are on a fairly safe basis; for the wages in these counties of the Northern District remain in nearly the same order through the last fifty years. At the same time it is easily seen that these wages are not so accurately known as those not interpolated, and it is well to notice in arguments based on such figures, to what extent the interpolated figures are involved.

A process very similar to that just employed is used in giving marks at school to students who are absent from a lesson; attention is paid both to the particular student's general place in the class order, and to the average value of the marks obtained by the rest of the class in the lesson missed.

Though the method be fairly complete it is very important to notice that interpolated figures rest on quite a different class

of evidence to those which are the result of direct evidence.
In some cases they may represent quantities
which have no existence (as in the case of school
marks) and which are only used for convenience of
calculation. In others they are simply figures
adopted as those which in default of definite knowledge appear
most probable. They must always be clearly indicated as inter-
polations; it is always well to state the method by which they
are obtained, and any subsidiary information which may be
regarded as direct evidence of their accuracy, and if practicable
they may be given not as exact, but as lying between certain
limits; thus the interpolated figures for Cheshire might be
written *12s. 6d.* to *13s. 6d.*, instead of *13s. 1d.*

Necessity for distinguishing interpolated figures.

Several different cases are met with in interpolation, some
of which are treated algebraically in the next section, while
others can be illustrated at once by numerical examples.

THE GRAPHIC METHOD.—If we know the values of quan-
tities at isolated positions, such as the numbers of the popula-
tion at the ages 25 to 35, 35 to 45, etc.; the
population in 1871, 1881, 1891, etc.; wages in
1860, 1866, 1870, etc.; the numbers whose wages are from
15s. to 20s., 20s. to 25s., etc., we may represent the facts by
such a diagram as—

Graphic method.
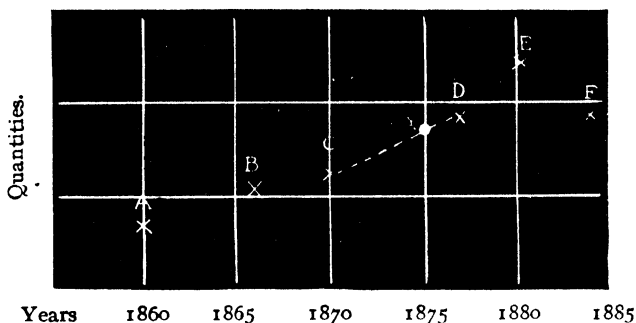
Suppose that we need the value of the quantity in 1875.
If we were only given the two points C and D, the simplest
hypothesis, and the one to be made in the absence of any
evidence to the contrary, is that the quantity increased
uniformly between C and D; representing such an increase
by the straight line C D, the height of the point *x* will represent
the quantity in 1875.

If the point E is also given, the hypothesis represented by the straight lines C D, D E will not stand, for it assumes a sudden break in the regularity at the point D in 1877, for which there is no evidence. We must take into account all the points given, and through them all a line must be drawn whose curvature is as smooth as possible, for in the absence of evidence' to the contrary, sudden changes in the quantities may be assumed not to exist. Such a curve can be constructed on mathematical principles, or may be drawn freehand; if the latter, it will often be quite as near the facts as the arguments will allow us to go.

This method only applies to continuous quantities, such as numbers at different ages, population at different dates, earners at different wages in a very large group of wages. Thus for all England the average wage must change gradually, but the wage of the London builders changed suddenly as the result of strikes and arrangements at certain dates. In this case we must draw the figure to correspond as closely as possible to the evidence, such as—
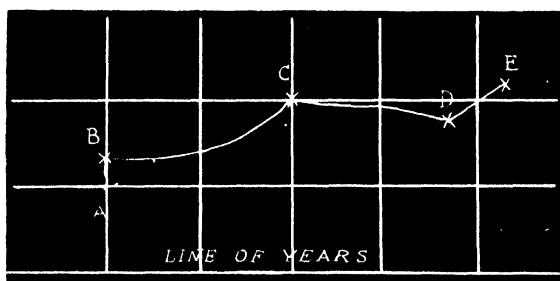
where A B represents a sudden rise; B C a gradually accelerated increase due to improving trade, C D a slow falling off from the wage reached at C, and D E a determined and successful effort to recover the lost ground.

PERIODIC FIGURES.—If we know the annual averages of figures which have a yearly period and a sufficient number of monthly averages to estimate the periodic fluctuations by the method described on pp. 160 *seq.*, we can interpolate figures for any month for which the returns are incomplete with fair accuracy. Thus if we are dealing with the numbers of unemployed as given in the *Labour Gazette*, we find a periodicity which is not very strongly marked in all the months, but there

is in general a fall in the spring and a rise in the late autumn, and June is generally the minimum month. We can then make use of the small diagrams on pp. 165–6, and, having marked in all the information we have, draw the waves on the rising, stationary, or descending line of averages, so that the fluctuating lines shall pass through all the given points. We can obtain an idea of the accuracy of the resulting figures by noticing the general characteristics of the given figures; we find that the percentage unemployed has never changed more than two units in one month, that there are no fluctuations which have lasted less than three or four months, and that the percentages have never been below 1 or above 10. Finally, we can look at the trade history of particular dates, and in the light we thus obtain reject any improbable figures.

USE OF SUBSIDIARY CURVES.—If we are able, by the methods described in Chapter VII, p. 158 or p. 174, to find a close connection between two series, we can use the more complete of them to assist the interpolation of any missing figures in the other. We must first investigate carefully the closeness and nature of correspondence at the dates for which we have complete figures in both series. Then we can draw diagrams, similar to those facing p. 155, one of the lines being incomplete. Then completing the broken line, so as to bring it into as close resemblance with the completed line as the given points allow, we shall obtain the most probable values for the missing figures. The accuracy of the result can be tested as in the previous case. This method may reasonably be used in interpolating figures for the yield from one source of revenue by means of the yield from another; for the value of exports from that of imports; for the marriage rate from foreign trade; for the wages in one district from those in another; for the number of unemployed from the changes in consumption of foods; for changes in parts of the population, when we know the changes in the whole, and for many other series.

SECTION 2.—ALGEBRAIC TREATMENT.

The problem of interpolation to which most attention has been given may be stated as follows :—When one quantity is subject to continuous regular change, and a second quantity changes in connection with it, and we know or can estimate

directly only some discontinuous values of this second quantity, it is required to find the probable values of the second quantity which correspond to given values of the first : for instance, given the expectation of life at the ages 15, 20, 25, etc., it is required to find it for intermediate ages; given the population of the country in 1871, 1881, 1891, 1901, find it at intermediate dates. The only permissible assumptions are that the quantity changes continuously, that is with no break at any figure, and that the rate of change of the quantity is also continuous, that is that the line representing its value is not angular, but smooth. The problem can only be attacked systematically by the use of the algebraic method of finite differences, and it is necessary to begin with definitions of notation and to obtain certain fundamental formulæ.

1. Let y be a continuous function of x, and let $y_0, y_1, y_2 \ldots$ be the values of y when $x = x_0, x_1, x_2 \ldots$.

Arrange a table thus—

Values of x.	Values of y.	First Differences.	Second Differences.	Third Differences.
x_0	y_0			
		Δ_0^1		
x_1	y_1		Δ_0^2	
		Δ_1^1		Δ_0^3
x_2	y_2		Δ_1^2	
		Δ_2^1		Δ_1^3
x_3	y_3		Δ_2^2	
		Δ_3^1		Δ_2^3
x_4	y_4		Δ_2^3	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Here each Δ is obtained by subtracting the entry just higher than it in the previous column from that just lower than it; *e.g.*, $\Delta_0^1 = y_1 - y_0$, $\Delta_1^1 = y_2 - y_1$, $\ldots$ $\Delta_0^2 = \Delta_1^1 - \Delta_0^1$, $\ldots$ $\Delta_0^3 = \Delta_1^2 - \Delta_0^2 \ldots$ The table may be supposed to continue indefinitely downwards and to the right.

We have at once—

$$\Delta_0^2 = \Delta_1^1 - \Delta_0^1 = (y_2 - y_1) - (y_1 - y_0) = y_2 - 2y_1 + y_0$$
$$\Delta_t^2 = y_{2+t} - 2y_{1+t} + y_t, \text{ where } t \text{ is any integer,}$$
$$\Delta_0^3 = (y_3 - 2y_2 + y_1) - (y_2 - 2y_1 + y_0) = y_3 - 3y_2 + 3y_1 - y_0$$
$$\Delta_t^3 = y_{3+t} - 3y_{2+t} + 3y_{1+t} - y_t$$

and generally, by an induction similar to that commonly used in the proof of the Binomial Theorem and involving the same coefficients—

$$\Delta_0^r = y_r - r \cdot y_{r-1} + \frac{r(r-1)}{1 \cdot 2} y_{r-2} - \frac{r(r-1)(r-2)}{1 \cdot 2 \cdot 3} y_{r-3} + \ldots$$
$$\text{to } \overline{r+1} \text{ terms} \ldots \ldots \ldots (a)$$

and $\quad \Delta_r^r = y_{r+t} - r \cdot y_{r+t-1} + \dfrac{r(r-1)}{1 \cdot 2} y_{r+t-2} + \ldots$ to $\overline{r+1}$ terms, $\quad (\beta)$

where r is any integer.

We have also—

$y_1 = y_0 + \Delta_0^1$, and $y_2 = y_1 + \Delta_1^1 = (y_0 + \Delta_0^1) + (\Delta_0^1 + \Delta_0^2) = y_0 + 2\Delta_0^1 + \Delta_0^2$,
and similarly $\quad \Delta_1^1 = \Delta_0^1 + \Delta_0^2$,

and $\quad \Delta_2^1 = \Delta_1^1 + \Delta_1^2 = (\Delta_0^1 + \Delta_0^2) + (\Delta_0^2 + \Delta_0^3) = \Delta_0^1 + 2\Delta_0^2 + \Delta_0^3$.

$$\therefore \; y_3 = y_2 + \Delta_2^1 = y_0 + 3\Delta_0^1 + 3\Delta_0^2 + \Delta_0^3,$$

and similarly $\quad \Delta_3^1 = \Delta_0^1 + 3\Delta_0^2 + 3\Delta_0^3 + \Delta_0^4$.

Continuing this process we again have the Binomial Coefficients, so that—

$$y_r = y_0 + r \cdot \Delta_0^1 + \frac{r(r-1)}{1 \cdot 2} \Delta_0^2 + \ldots \text{ to } \overline{r+1} \text{ terms } \ldots \quad (\gamma)$$

$$\Delta_r^t = \Delta_0^t + r \cdot \Delta_0^{t+1} + \frac{r(r-1)}{1 \cdot 2} \Delta_0^{t+2} + \ldots \text{ to } \overline{r+1} \text{ terms } \quad (\delta)$$

and starting further down the scale—

$$y_{r+s} = y_s + r \cdot \Delta_s^1 + \frac{r(r-1)}{1 \cdot 2} \Delta_s^2 + \ldots \text{ to } \overline{r+1} \text{ terms, } \ldots \quad (\epsilon)$$

where s is any integer.

For example, let $y = x^4$, and let the values of x be $0, h, 2h, 3h \ldots$

Values of x.	Values of y.			Differences.		
		First.	Second.	Third.	Fourth.	Fifth.
0	0					
		h^4				
h	h^4		$14h^4$			
		$15h^4$		$36h^4$		
$2h$	$16h^4$		$50h^4$		$24h^4$	
		$65h^4$		$60h^4$		0
$3h$	$81h^4$		$110h^4$		$24h^4$	
		$175h^4$		$84h^4$		0
$4h$	$256h^4$		$194h^4$		$24h^4$	
		$369h^4$		$108h^4$		
$5h$	$625h^4$		$302h^4$			
		$671h^4$				
$6h$	$1296h^4$					

Formula (α) gives $\Delta_0^4 = (256 - 4 \times 81 + 6 \times 16 - 4 \times 1 + 0)h^4$
$\qquad\qquad\qquad\qquad = 24h^4$, where r is taken as 4.

Formula (β) gives $\Delta_2^5 = (7^4 - 5 \times 6^4 + 10 \times 5^4 - 10 \times 4^4 + 5 \times 3^4 - 2^4)h^4 = 0$, where $r = 5$, $t = 2$.

Formula (γ) gives $(5h)^4 = (0 + 5 + 10 \times 14 + 10 \times 36 + 5 \times 24 + 0)h^4 = 625h^4$, where $r = 5$.

Formula (δ) gives $\Delta_2^3 = (36 + 2 \times 24 + 0)h^4 = 84h^4$,
$\qquad\qquad\qquad\qquad\qquad$ where $r = 2$, $t = 3$,

and—

Formula (ϵ) gives $(5h)^4 = (16 + 3 \times 65 + 3 \times 110 + 84)h^4$
$\qquad\qquad\qquad\qquad = 625h^4$, where $r = 3$, $s = 2$.

2. If the relation between y and x is of the form

$$y = a_0 + a_1x + a_2x^2 + \ldots + a_nx^n,$$

and the values of x are in Arithmetic Progression, viz. $x_0, \overline{x_0 + h},$ $\ldots \overline{x_0 + (n-1)h}$, then it can be shown that $\triangle_0{}^n = a_n \cdot h^n\underline{\lfloor n}$, and that there are no higher differences.

For $\Delta_0{}^1 = a_0 - a_0 + a_1(x_0 + h - x_0) + \ldots + a_n\{(x_0 + h)^n - x_0{}^n\}$
$\qquad = ha_1 + \ldots + a_n\{nhx_0{}^{n-1} + \text{lower powers of } x_0\},$
$\quad \Delta_1{}^1 = ha_1 + \ldots + a_n\{nh(x_0 + h)^{n-1} + \text{lower powers of } \overline{x_0 + h}\}$
$\quad \Delta_0{}^2 = 2h^2a_2 + \ldots + a_n\{n(n-1)h^2x_0{}^{n-2} + \text{lower powers of } x_0\}.$

Thus $\Delta_0{}^1$, $\Delta_0{}^2$ contain no higher powers than $x_0{}^{n-1}$ and $x_0{}^{n-2}$ respectively.

Continuing this process—

$$\Delta_0{}^n = a_n n(n-1) \ldots 3 \cdot 2 \cdot 1 h^n = a_n h^n n! \quad \ldots \quad (\zeta)$$

and $\Delta_0{}^{n+1}$ and higher differences disappear.

In the example above where—

$$y = x^4, \ a_n = 1, \ n = 4, \ \Delta_0{}^4 = 1 \cdot h^4 \cdot 4! = 24h^4, \text{ and } \Delta_0{}^5 = 0.$$

Conversely if we assume that there is no difference above the nth, it is shown in the following note that the equation between y and x is of the form $y = a_0 + a_1x + \ldots + a_nx^n$.

Note.—The relation between Differences and Derived Functions (or Differential Coefficients) is very important in the theory of the former, and can be exhibited concisely by the method of operators.

Using the usual notation of the calculus, we have by Taylor's Theorem—

$f(x+h) = f(x) + hf'(x) + \frac{1}{2!}h^2f''(x) + \ldots = e^{hD} \cdot f(x)$, where D stands for the

operation of differentiation, and e^{hD} is to be expanded into $1 + hD + \frac{1}{2!}h^2D^2 + \ldots$ and then applied term by term to $f(x)$. The use of D as an algebraic symbol is justified because of the relationships $D\{Df(x)\} = D^2f(x)$, $D^m\{D^nf(x)\} = D^{m+n}f(x)$, $aD(f(x)) = D(af(x))$, etc.

Now $\Delta f(x) = f(x+h) - f(x) = (e^{hD} - 1)f(x).$

$\Delta\{af(x)\} = a\Delta f(x)$, $\Delta\{\Delta f(x)\} = \Delta^2f(x)$, $\Delta^m(\Delta^nf(x)) = \Delta^{m+n}f(x)$, and Δ can be used as an algebraic symbol.

Hence $\Delta^n \equiv e^{hD} - 1$

$$\Delta^n \equiv (e^{hD} - 1)^n = (hD + \frac{1}{2!}h^2D^2 + \ldots)^n = h^nD^n(1 + \tfrac{1}{2}hD + \tfrac{1}{6}h^2D^2 + \ldots)^n$$

$$= h^nD^n(1 + \frac{n}{2}hD + \frac{n(3n+1)}{24}h^2D^2 + \ldots) \quad \ldots \ldots \quad (i)$$

and $hD \equiv \log(1 + \Delta)$

$h^nD^n \equiv \{\log(1 + \Delta)\}^n = (\Delta - \tfrac{1}{2}\Delta^2 + \tfrac{1}{3}\Delta^3 - \ldots)^n$

$$= \Delta^n(1 - \frac{n}{2}\Delta + n\frac{(3n+5)}{24}\Delta^2 + \ldots) \quad \ldots \ldots \quad \ldots \ldots \quad (ii)$$

Now if $f(x) = a_0 + a_1x + \ldots + a_nx^n$, $D^nf(x) = a_n \cdot n!$, and $D^{n+1}f(x) = 0$ $= D^{n+2}f(x) \ldots$

$\therefore \Delta^nf(x) = h^na_nn!$, and $\Delta^{n+1}f(x) = h^{n+1}D^{n+1}(1 + \ldots)f(x) = 0$ from equation (i) as in the text.

Conversely if $\Delta^{n+1}f(x) = 0 = \Delta^{n+2}f(x) = \ldots$, then from (ii) $D^{n+1}f(x)$

$= 0$, $D^n f(x) = \text{const} = c_n$, $D^{n-1}f(x) = c_n x + c_{n-1}$, $D^{n-2}f(x) = \frac{1}{2} c_n x^2 + c_{n-1} x + c_{n-2}$,

and $f(x) = \frac{1}{n!} c_n x^n + \ldots + c_1 x + c_0$.

Hence if the nth difference is constant, the function is rational, integral, and of the nth degree.

Newton's interpolation formula, discussed below (κ), can quickly be obtained by the use of operators; thus—

$$y = f(x_0 + h) = e^{hD}f(x_0) = (1+\Delta)^{\frac{h}{h}}f(x_0), \text{ since } e^{hD} = 1 + \Delta,$$
$$= f(x_0) + \frac{k}{h}\Delta f(x_0) + \frac{1}{2} \cdot \frac{k}{h}\left(\frac{k}{h} - 1\right) \cdot \Delta^2 f(x_0) + \ldots$$
$$= y_0 + \frac{x - x_0}{h}\Delta_0{}^1 + \frac{x - x_0}{h} \cdot \frac{x - x_0 - h}{2h}\Delta_0{}^2 + \ldots, \text{ where } x = x_0 + h.$$

When the nth difference (or the nth derived function) is zero, formula β shows that

$$y_{n+t} - n y_{n-1+t} + \frac{n(n-1)}{1 \times 2} y_{n-2+t} \ldots \pm y_t = 0 \quad \ldots \quad (\eta)$$

for all values of t.

3. The common formula of interpolation depends on the assumption that a continuous function, $y = f(x)$, can represent the observations in the neighbourhood of the positions for which values are to be found.

It is assumed that the function can be expanded in powers of x, as is generally the case with continuous functions,* we may write—

$$y = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n, \quad \ldots \quad \ldots \quad (\theta)$$

where n, the index of the highest power of x, is still to be decided. By proper choice of a_0, $a_1 \ldots a_n$ this equation can be satisfied by any $(n+1)$ pairs of values of (x and y). Thus for the straight line $y = a_0 + a_1 x$, two points (or pairs of values) can be chosen, for the parabola $y = a_0 + a_1 x + a_2 x^2$ three points, and so on.

The simplest form is $y = a_0 + a_1 x$, and the use of this assumes that interpolation by proportional parts (the method generally employed in using logarithmic, trigonometric and other mathematical tables) is sufficiently accurate. In this case the first difference and the first derived function (or gradient) are constant.

The parabola takes account of three values, and its use assumes a uniform change of gradient, the second difference and the second derived function being constant.

The introduction of further terms allows for variation of

---

* More exactly for functions which are continuous, and whose derived functions of all orders are continuous, and not infinite, at the values of x in question.

higher differences, and the closing the expansion at the nth term corresponds to constancy of the nth difference.

If the problem is to interpolate in a known mathematical function we can test how far the neglect of the variation in the nth difference can affect the calculation. Thus in the 7-figure logarithm table we have—

Number.	Logarithm.	First.	Second.	Third.	Fourth.	Fifth.
20	1.3010300	.0211893				
21	1.3222193	.0202034	−.0000859			
22	1.3424227	.0193051	−.0008983	+.0000876	−.0000110	
23	1.3617278	.0184834	−.0008217	+.0000766	−.0000095	+.0000015
24	1.3802112	.0177288	−.0007546	+.0000671	−.0000080	+.0000015
25	1.3979400	.0170333	−.0006955	+.0000591	−.0000064	+.0000016
26	1.4149733	.0163905	−.0006428	+.0000527		
27	1.4313638					

Here the successive differences diminish regularly and the sixth difference is not greater than ·0000001.

In applications to statistics we do not in general know the function and we have to assume that it exists and can be expanded in a series whose convergence is sufficiently rapid to allow us to neglect all terms after, say, the fifth, or, put less accurately, we assume that the causes which produce the totals have effects which change gradually from point to point, so that the variation of these changes is but slight over a small region.

4. Let $y_0, y_1 \ldots y_n$ be the values of y which correspond to equally spaced values of x, viz. $x_0, x_0 + h, x_0 + 2h \ldots x_0 + nh$. Then the coefficients in equation (θ) can be determined, but the arithmetic work is very arduous, and a more useful form is obtained in terms of differences.

Consider the equation—

$$y = y_0 + \frac{x - x_0}{h}\Delta_0{}^1 + \frac{x - x_0}{h} \cdot \frac{x - x_0 - h}{2h}\Delta_0{}^2 + \frac{x - x_0}{h} \cdot \frac{x - x_0 - h}{2h} \cdot$$

$$\frac{x - x_0 - 2h}{3h}\Delta_0{}^3 + \ldots \text{ to } n+1 \text{ terms} \quad \ldots \quad (\kappa)$$

(Newton's formula)

If $x = x_0$, $y = y_0$.

If $x = x_0 + h$, $y = y_0 + \Delta_0{}^1 = y_1$.

If $x = x_0 + 2h$, $y = y_0 + 2\Delta_0{}^1 + \Delta_0{}^2 = y_2$.

If $x = x_0 + rh$, $y = y_0 + r \cdot \Delta_0{}^1 + \frac{r(r-1)}{1 \cdot 2}\Delta_0{}^2 +$ to $r+1$ terms, the subsequent terms vanishing, and therefore by equation (γ), $y = y_r$.

Hence (κ), which is easily seen to be of the nth degree, is satisfied by the n pairs of values in question.

E.g., to find $y = \log 20\cdot5$ from the table above.

$x_0 = 20$, $h = 1$, $x = x_0 + \cdot5$, $y_0 = 1\cdot3010300$, $\Delta_0{}^1 = \cdot0211893$, etc.

$y = 1\cdot3010300 + \cdot5$ of $\cdot0211893 + \frac{1}{2}(\cdot5)(-\cdot5)(-\cdot0009859) + \frac{1}{6}(\cdot5)(-\cdot5)$
$(-1\cdot5)$ of $\cdot0000876 + \frac{1}{24}(\cdot5)(-\cdot5)(-1\cdot5)(-2\cdot5)(-\cdot0000110)$
$+ \frac{1}{120}(\cdot5)(-\cdot5)(-1\cdot5)(-2\cdot5)(-3\cdot5)$ of $\cdot0000015$.

Using the first two terms, we have $y = 1\cdot3116247$.

,,	,,	three	,,	,,	$1\cdot3117479$.
,,	,,	four	,,	,,	1.3117534.
,,	,,	five	,,	,,	$1\cdot3117538$.
,,	all terms		,,	,,	$1\cdot3117538$.

The true value is $1\cdot3117539$.

Applications to statistical data are given below, p. 233.

5. Conversely, if we know y, we have an equation for x, which can be solved by Horner's method or otherwise.

Thus to determine *the median* using four observations we may proceed as follows. Let there be y_0, y_1, y_2, y_3 persons whose wages are less than x_0, $x_0 + h$, $x_0 + 2h$, $x_0 + 3h$ units respectively, and let there be $(2y_m - 1)$ persons all together, so that the value of x, x_m, corresponding to y_m is the median.

Then
$$y_m = y_0 + \frac{x_m - x_0}{h}\Delta_0{}^1 + \frac{x_m - x_0}{h} \cdot \frac{x_m - x_0 - h}{2h}\Delta_0{}^2$$
$$+ \frac{x_m - x_0}{h} \cdot \frac{x_m - x_0 - h}{2h} \cdot \frac{x_m - x_0 - 2h}{3h}\Delta_0{}^3,$$

a cubic equation to determine x_m.

We are free, of course, to take x_0 as the beginning of any grade we please, and it should be so chosen that the median is in the central grade included in the interpolation. Thus if we use the cubic equation just written the grade $x_0 + h$ to $x_0 + 2h$ should be that containing the median.

The formula on p. 107 (2) is obtained by neglecting the 2nd and higher differences, and taking the grade x_0 to $x_0 + h$ to include the median. Then $y_m = y_0 + \frac{x_m - x_0}{h}(y_1 - y_0)$, and therefore—

$$x_m = x_0 + \frac{y_m - y_0}{y_1 - y_0} \cdot h.$$

To find the *mode* we again take y as the cumulative number

up to the value x. It is found that it is simplest and generally sufficient to depend on four observations, such that the mode is between the second and the third. We then use the first four terms of equation (κ) and find for what value of x the curve is steepest and therefore the number of cases per unit of the abscissa is greatest. $D_x y$ is to be a maximum, and therefore $D_x^2 y$ zero.

$$0 = D_x^2 y = \frac{1}{h^2}\Delta_0{}^2 + \frac{x - x_0 - h}{h^3}\Delta_0{}^3.$$

Hence $\quad x = x_0 + h - \dfrac{h\Delta_0{}^2}{\Delta_0{}^3} = x_0 + h + \dfrac{(u_2 - u_1)h}{(u_3 - u_1) + (u_2 - u_3)},$

where u_1, u_2, u_3 are written for $y_1 - y_0$, $y_2 - y_1$, $y_3 - y_2$ and are the number of cases between x_0 and $x_0 + h$, $x_0 + h$ and $x_0 + 2h$, and $x_0 + 2h$ and $x_0 + 3h$ respectively. If the mode is in the second grade $u_2 > u_1$ and $u_2 > u_3$. The formula shows how the interval $x_0 + h$ to $x_0 + 2h$ is to be divided to obtain the position of the mode (see p. 100).

Here the fourth differences of the y's, that is the third differences of the u's, are neglected.

6. *Central Differences.*—In interpolation we generally have to depend on those values of y with regard to which the region where we wish to ascertain values is centrally situated, and formula (θ) is in some respects awkward for that purpose. Equivalent formulæ, which avoid the want of symmetry, have been devised, in which so-called " central differences " are used. No new principle is involved, for these formulæ are obtainable by transformation from (θ). The differences hitherto used may be distinguished as " ascending differences."

A suitable notation is as follows:—

$x_{-2} = x_0 - 2h$	y_{-2}				
		$\delta_{-\frac{3}{2}}$			
$x_{-1} = x_0 - h$	y_{-1}		δ^2_{-1}		
		$\delta_{-\frac{1}{2}}$		$\delta^3_{-\frac{1}{2}}$	
x_0	y_0		δ^2_0		δ^4_0
		$\delta_{\frac{1}{2}}$		$\delta^3_{+\frac{1}{2}}$	
$x_1 = x_0 + h$	y_1		δ^2_1		δ^4_1
		$\delta_{\frac{3}{2}}$		$\delta^3_{+\frac{3}{2}}$	
$x_2 = x_0 + 2h$	y_2		δ^2_2		
		$\delta_{\frac{5}{2}}$			
$x_3 = x_0 + 3h$	y_3				

Here $\delta_{\frac{1}{2}} = y_1 - y_0$; $\delta^2_0 = \delta_{\frac{1}{2}} - \delta_{-\frac{1}{2}} = y_1 - 2y_0 + y_{-1}$; $\delta_0{}^4 = y_2 - 4y_1 + 6y_0 - 4y_{-1} + y_{-2}$, etc.

Let the value of x for which a value of y is to be found divide the interval x_0 to x_1 in the ratio $p : q$, so that $x = x_0 + ph = x_1 - qh$ and $p + q = 1$.

Then it will be found by substitution that the formula—

$$y = py_1 + qy_0 - \tfrac{1}{6}pq\{(p+1)\delta^2_1 + (q+1)\delta_0^2\}$$
$$+ \tfrac{1}{120}pq(p+1)(q+1)\{(p+2)\delta^4_1 + (q+2)\delta_0^4\}, \ . \ . \ . \ (\lambda)$$

which (by writing $q = 1 - p$) is seen to be a rational integral function of the 5th degree in p, and similarly in q, is satisfied by the six pairs of values $(x_{-2}\ y_{-2})\ (x_{-1}\ y_{-1})\ .\ .\ .\ (x_3\ y_3)$; while, if the term involving the 4th differences is omitted, the four pairs $(x_{-1}\ y_{-1})$ . . . $(x_2\ y_2)$ satisfy it.

As an illustration of the notation we may write $y_1 = \log 23$ in the table on p. 226, and taking $p = \cdot 2$ calculate $\log 22 \cdot 2$.

Log $22 \cdot 2 = \cdot 2 \log 23 + \cdot 8 \log 22$

$$- \frac{\cdot 16}{6}\{1 \cdot 2 \text{ of } (-\cdot 0008217) + 1 \cdot 8 \text{ of } (-\cdot 0008983)\}$$

$$+ \frac{\cdot 16 \times 1 \cdot 2 \times 1 \cdot 8}{120}\{2 \cdot 2 \text{ of } (-\cdot 000095) + 2 \cdot 8 \text{ of } (-\cdot 0000110)\}$$

$$= 1 \cdot 3462837 + \cdot 0000694 - \cdot 0000002 = 1 \cdot 3463529.$$

The true value is $1 \cdot 3463530$.

The importance of the formula is, however, more apparent when we have no general algebraic function, but wish to interpolate from neighbouring values only.

7. *Lagrange's Formula.*—The formulæ $(\zeta)\ (\eta)\ (\theta)\ (\kappa)$ and (λ) all relate to the case where the observed values of x are equidistant each from the next. There is no such simple method of interpolation where the distances are not equal. An equation is given by Lagrange which is of the nth degree and satisfied the $\overline{n+1}$ pairs of values $(x_0\ y_0),\ (x_1\ y_1)\ .\ .\ .\ (x_n\ y_n)$ whatever the relation between the x's may be, and it may be written as follows:—

$$y = y_0 \frac{(x-x_1)(x-x_2)\ .\ .\ .\ (x-x_n)}{(x_0-x_1)(x_0-x_2)\ .\ .\ .\ (x_0-x_n)} + y_1 \frac{(x-x_0)(x-x_2)\ .\ .\ .\ (x-x_n)}{(x_1-x_0)(x_1-x_2)\ .\ .\ .\ (x_1-x_n)}$$
$$+ \ .\ .\ .\ + y_n \frac{(x-x_0)(x-x_1)\ .\ .\ .\ (x-x_{n-1})}{(x_n-x_0)(x_n-x_1)\ .\ .\ .\ (x_n-x_{n-1})}\ .\ .\ (\mu)$$

The numerator in any fraction, say the multiplier of y_t, is obtained by multiplying the factors $(x - x_0)\ (x - x_1)\ .\ .\ .$ $(x - x_n)$ omitting $x - x_t$; the denominator is obtained from the numerator by writing x_t for x.

It is evident that when $x = x_t$ every fraction is zero except the multiplier of y_t, which is unity, and therefore $y = y_t$.

8. We may now reconsider the assumptions made when we took equation (θ) to express the relation between y and x.

If y and x are connected by any functional law, that is if y is determinate for all given values of x, without which assumption most problems of interpolation are meaningless, then y can be expressed as a function of x, say $y = f(x)$. If the function and its derivatives are continuous then by Maclaurin's Theorem—

$$y = f(o) + xf^1(o) + \frac{x^2}{2!}f^2(o) + \frac{x^3}{3!}f^3(o) + \ldots \text{ continued indefinitely.}$$

If $f^{n+1}(o)$ and following coefficients are very small, and x is never large, the terms from the $\overline{n+2}^{\mathrm{nd}}$ onwards become negligible in comparison with earlier terms, so that the first $\overline{n+1}$ terms determine the value of y approximately. Now by the equations (i) and (ii), p. 224, f^{n+1} is small when $\triangle^{n+1}$, $\triangle^{n+2}$, . . . are small, and *vice versa*. Hence we have the following general statement : any functional relation between y and x reduces to the parabolic equation of the nth degree (θ), if the differences of orders higher than the nth vanish, and if these differences do not vanish but are small, equation (θ) is still an approximate expression for the relation.

Now if the line drawn through the given points is to have continuous and slowly changing curvature, it is easily verified that the second differences for points near together are not large, for a rapid change in the rate of increase of the ordinate means a rapid change of curvature; and if we construct a second curve with the same abscissæ and the first differences as ordinates, small third differences will indicate absence of rapid change in the first, and so on; but beyond this point it is not easy to see the connection between the hypothesis underlying interpolation and the diminution of successive differences. The converse, however, is clearer; if in any series of figures it is found experimentally that the successive differences tend to disappear, then any curve which passes through the points is expressed approximately by the parabolic equation. De Morgan states this conclusion thus :—
" If we take n points near each other, and having their abscissæ in arithmetic progression, with a small or at least not very large common difference, and their ordinates not very unequal . . . the parabola of the $\overline{n-1}^{\mathrm{th}}$ order will very nearly coincide

with any regular curve of the same general appearance, at least between the same points." Boole's explanation is :— " It is customary to assume for the general expression of the values under consideration a rational and integral function of x, and to determine the constants by the given conditions. This assumption rests upon the supposition (a supposition, however, actually verified in the case of all tabulated functions*) that the successive orders of differences rapidly diminish."

Since, from equation (i), p. 224, when h is small, the influences of the successive differences for any curve are smaller as their order becomes higher, it is a legitimate process to build up a series of values of any function on the hypothesis that the higher differences vanish.

If a freehand curve is drawn so as to pass through the chosen fixed points, and to have curvature which changes as slowly as possible, a line will be obtained which lies very near that given by equation (θ). Such a line would be similar to the track of a bicyclist who was riding so as to pass over several marks, or just to avoid several obstacles.

9. It is clear from the above analysis that we can make a smooth continuous curve pass through any number of points we please; for with the parabolic equation (θ) there are never any sudden jumps in the values of y, $\dfrac{dy}{dx}$ or $\dfrac{d^2y}{dx^2}$, as x changes continuously; and we can obtain as many linear equations (which have always real values) as there are constants, simply by taking n in the original equation to be the number of fixed points.

If we have, let us say, 10 points, as—

---
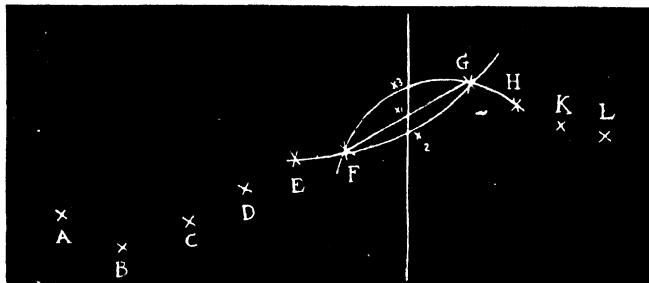
* That is mathematical functions such as $\displaystyle\int_{0}^{x} e^{-x^2}\, dx$, not statistical approximations. ⚬

and wish to find a point on a fixed vertical line between F and G, we can either take only F and G into consideration, and, joining them by a straight line, obtain the point x_1; or considering E, F, and G, or F, G, and H, draw parabolas and obtain x_2 or x_3; or considering E, F, G, and H, draw a parabola of the third order, which would have a point of inflexion near F; this would be approximately the path a bicyclist might follow if he had to start from E, and ride to a near point H, passing close to F and G. If we now include D and K (if our bicyclist has to start from D, pass E, F, G, and H, and reach K) we shall modify the curvature throughout; and as we include more and more points shall continue to affect slightly the path F G. If the inclusion of the nearer points tends to make the line F G approximate more and more closely to a final position, while the further inclusion of the more distant points throws it further away, we may conclude that the positions of these further points are not governed by the same numerical conditions as the nearer one. Thus in a " table of survivals " the figures for ages under 5 years are not distributed in accordance with the curve determined by the figures for higher ages; in a table showing wages, it may be seen that those of highly paid workmen are not governed by the same causes as those lower in the scale. On the other hand, the number in each census is dependent on all the previous numbers for more than one generation. In interpolating for the population of 1876 we shall obtain different figures according as we include 1851, '61, '71, '81, '91 only, or 1901 as well; and this is not surprising, for a mistake made in 1876 may not come to light till we have watched the growth of the population for twenty-five years. It is clear that the points far from the period in which the interpolation is to be done cannot be allowed so much influence as those nearer, and it appears experimentally that this condition is fulfilled in the method discussed; also, in series (κ) the successive coefficients begin to diminish with the rth term where $x < x_0 + (2r - 3)h$, that is with the co-efficient of the first difference when x is between x_0 and $x_0 + h$. It may be noticed that the wanderings of the curve are limited by the condition that a curve of the $n - 1^{th}$ order cannot have more than $\overline{n - 3}$ points of inflexion, for $\dfrac{d^2 y}{dx^2}$ has no term of a higher degree than x^{n-3}.

In the above illustration the intermediate points from F to G might be found from the five points D, E, F, G, H, 'or from E, F, G, H, K. These two curves may be welded together between F and G. 'The points near F are more accurately determined by the first, of 'which it is the middle; those near 'G by the second. The welding line should touch the first at F, the second at G. This is conveniently done by the use of the 'sine curve. This method is employed, I believe, at the Registrar-General's office.

It cannot be said that the present theory of statistical interpolation rests on an altogether satisfactory basis.* The principles which govern it are not well defined, and the mathematical analysis of the methods, by which the principles should be brought into relation with the facts, is incomplete. Yet it is perhaps unnecessary to labour after more refined methods, for interpolation cannot be precise unless we actually know the algebraic expression of the laws which govern the figures, and the method here discussed is found to satisfy the conditions empirically, while further refinements could only introduce slight modifications.

10. *Examples showing the Numerical Use of the Formulæ.*— (1) Given the number of wage-earners earning sums in 5s. groups, to estimate the number earning as much as 24s. and not so much as 25s.

		*Numbers per 1,000 Wage-Earners (Adult males)	DIFFERENCES.			
			1st.	2nd.	3rd.	4th.
Earning as much as 10s.	15s.	39				
			257			
	20s.	296		46		
			303		-144	
	25s.	599		-98		151
			205		7	
	30s.	804		-91		18
			114		25	
	35s.	918		-66		
			48			
	40s.	966				

(The leftmost bracket for rows 20s.–40s. is labelled "but not so much as")

* General Report on Wages, (C—6889; year 1893).

*. This remark does not apply to the interpolation in evaluating mathematical functions.

Neglect the increasing differences arising from the number earning less than 15s.

Using formula (κ), $x_0 = 20$ (shillings), $h = 5$, $y_0 = 296$, $\Delta_0{}^1 = 303$, $\Delta_0{}^2 = -98$, $\Delta_0{}^3 = 7$, $\Delta_0{}^4 = 18$.

At 25s., $y = 599$, from above table.

At 24s., $x = 24$, $y = 296 + \dfrac{4}{5}$ of $303 + \dfrac{4}{5} \cdot \dfrac{-1}{10}$ of $(-98) +$

$\dfrac{4}{5} \cdot \dfrac{-1}{10} \cdot \dfrac{-6}{15}$ of $7 + \dfrac{4}{5} \cdot \dfrac{-1}{10} \cdot \dfrac{-6}{15} \cdot \dfrac{-11}{20}$ of 18.

$= 296 + 242\cdot4 + 7\cdot84 + \cdot224 - \cdot3168 = 546$ (nearly).

The required number is therefore $599 - 546 = 51$.

Again at 23s., $x = x_0 + 3$, $y = 489$, and the number earning as much as 23s. and not so much as 24s. is 58.

(2) To make an estimate for the value of imports in the year 1813, the records for which were destroyed by fire.

Given value of imports in—

1810	-	-	- £39,202,000	-	-	- y_1.
1811	-	-	- 26,510,000	-	-	- y_2.
1812	-	-	- 26,163,000	-	-	- y_3.
1813	-	-	- ...	-	-	- y_4.
1814	-	-	- 33,755,000	-	-	- y_5.
1815	-	-	- 32,987,000	-	-	- y_6.
1816	-	-	- 27,431,000	-	-	- y_7.

From formulæ (η), using y_3 and y_5 only, and assuming that 2nd differences vanish,

$$y_5 - 2y_4 + y_3 = 0, \quad y_4 = 29,959.$$

From formulæ (η), using y_2 and y_6 as well, and assuming that 4th differences vanish,

$$y_6 + y_2 - 4(y_5 + y_3) + 6y_4 = 0, \quad y_4 = 30,029.$$

From formulæ (η), using y_1 and y_7 as well, and assuming that 6th differences vanish,

$$y_7 + y_1 - 6(y_6 + y_2) + 15(y_5 + y_3) - 20y_4 = 0, \quad y_4 = 30,421.$$

Here the first and second values are very near together, while the third differs; hence we adopt £30,000,000 as the value required.

(3) In Mr. Booth's *Life and Labour of the People*, *e.g.*, Vol. V, p. 46, a series of very useful diagrams is given showing the age distribution of various classes. The figures he uses are as follows :—

Ages						Proportion occupied per 10,000 of total aged 10-80.	Average at each year of age between given limits.
10–15 years	-	-	-	-	-	193·5	38·7
15–20 ,,	-	-	·	-	-	880	176
20–25 ,,	-	-	-	-	-	933	186·6
25–35 ,,	-	-	-	-	-	1636	163·6
35–45 ,,	-	-	-	-	-	1201	120·1
45–55 ,,	-	-	-	·	-	830	83
55–65 ,,	-	-	-	-	-	434	43·4
65–80 ,,	-	-	-	-	-	192·5	12·8

His diagram is drawn from the last column, the numbers in which form the ordinates for the middle of the corresponding age periods. The points so obtained are joined by straight lines. This method is sufficiently accurate for his purpose, but it will afford an interesting example of interpolation if we obtain some of the figures for intermediate years more closely.

Age.						Proportion occupied per 10,000 under x years.
$15 = x_1$	-	-	-	-	-	$193·5 = y_1$
$20 = x_2$	-	-	-	-	-	$1073·5 = y_2$
$25 = x_3$	-	-	-	-	-	$2006·5 = y_3$
$35 = x_4$	-	-	-	-	-	$3642·5 = y_4$
$45 = x_5$	-	-	-	-	-	$4843·5 = y_5$
$55 = x_6$	-	-	-	-	-	$5673·5 = y_6$
$65 = x_7$	-	-	-	-	-	$6107·5 = y_7$
$80 = x_8$	-	-	-	-	-	$6300 = y_8$

Use Lagrange's formula (μ) to determine the number under 30 years, ignoring persons over 55. Thus $x = 30$.

$$y = 193·5 \times \frac{10 \cdot 5(-5)(-15)(-25)}{(-5)(-10)(-20)(-30)(-40)}$$
$$+ 1073·5 \times \frac{15 \cdot 5(-5)(-15)(-25)}{5(-5)(-15)(-25)(-35)}$$
$$+ 2006·5 \times \frac{15 \cdot 10(-5)(-15)(-25)}{10 \cdot 5(-10)(-20)(-30)}$$
$$+ 3642·5 \times \frac{15 \cdot 10 \cdot 5(-15)(-25)}{20 \cdot 15 \cdot 10(-10)(-20)}$$
$$+ 4843·5 \times \frac{15 \cdot 10 \cdot 5(-5)(-25)}{30 \cdot 25 \cdot 20 \cdot 10(-10)}$$
$$+ 5673·5 \times \frac{15 \cdot 10 \cdot 5(-5)(-15)}{40 \cdot 35 \cdot 30 \cdot 20 \cdot 10}$$

$$= 2879.$$

Mr. Booth's diagram gives 2824·5 for the same position, using $y_3 + y_4$ only.

If in the formula the quantities y_2, y_3, y_4, y_5 only are used, y is found to be 2869.

Lagrange's formula as used above is equivalent to the assumption that the 6th differences vanish when the ages are uniformly graded. Write a, b, c for the values of y at 30, 40 and 50 years.

Using formula (β) or (η) for the values $y_1, y_2, y_3, a, y_4, b, y_5$ we have $y_1 - 6y_2 + 15y_3 - 20a + 15y_4 - 6b + y_5 = 0$, and similarly

$$y_2 - 6y_3 + 15a - 20y_4 + 15b - 6y_5 + c = 0$$
and $$y_3 - 6a + 15y_4 - 20b + 15y_5 - 6c + y_6 = 0.$$

Whence by straightforward solution $a = 2879$ as above. This method, when applicable, is simpler than Lagrange's formula.

(4) As an example of the determination of the median and the mode, we will use the figures already employed on p. 69, which may be retabulated thus :—

Earning less than	x.	y.		Differences.		
\$.25	—1	0				
			317			
·75	0	317		1157		
			1472		—1332	
1.25	1	1789		—175		
			1297		—152	
1.75	2	3086		—327		—137	+15
			970		—137	
2.25	3	4056		—464		
			506			
2.75	4	4562				

The whole number of persons is 5123. To find the median put $y = 2562$, and use the entries from $x = 0$ to $x = 4$.

Then $2562 = 317 + 1472x - \frac{1}{2}$ of $175x(x-1) - \frac{1}{6}$ of $152x(x-1)(x-2)$
$\qquad\qquad + \frac{1}{24}$ of $15x(x-1)(x-2)(x-3)$,
if we stop at the 4th difference.

$\therefore 61488 = 7608 + 36122x - 111x^2 - 698x^3 + 15x^4$, and the solution by Horner's method is $x = 1·5715$.

Hence the median is at \$.75 + 1·5715 of ·50 = \$1.536.

Another method is to suppose x expressed as a function of y,[*] and to write Lagrange's formula—

$$x = \frac{(y - y_1)(y - y_2)(y - y_3)}{(y_0 - y_1)(y_0 - y_2)(y_0 - y_3)} x_0 + \quad + \quad +.$$

---

[*] *Cf.* Edgeworth in the *Statistical Journal*, 1898, p. 698.

If we use four entries only in the above table, we have—

$$x = \frac{(2562-1789)(2562-3086)(2562-4056)}{-1472 \times -2769 \times -3739} \text{ of } o \quad + \quad + \quad +,$$

whence $x = 1.5624$ and the median is \$1.531.

This method is suitable for working on a calculating machine.

To find the *mode* use the entries from $x = -1$ to $x = 2$.

The second and third differences in the formula of p. 228 are now 1157 and -1332.

The required value is \$.75 $+ \dfrac{1157}{1332}$ of $.50 = 1.18.

Variations of method can be used, leading to slightly different results. The mode is, in fact, not precisely determinate when the grading is so wide and the higher differences do not tend to zero.

This method is applicable to such problems as the determination of the date at which the population, the marriage, birth, and death rates, etc., increased most rapidly; at what age the chance of death increases most, etc.[*]

11. An important group of problems of interpolation arise when the original returns have to be corrected, *e.g.*, the determination of the distribution by age from the census returns.

We have now the problem of drawing a smooth line in the neighbourhood of a great number of points, but not necessarily through any of them. The assumption is that the returns are insufficient in number or deficient in accuracy, and that they indicate a regular distribution which it is required to represent.

(1) One method is to assume that the averages over fairly large groups are accurate, and to these averages to apply any of the methods already discussed.

(2) A second method has been used in the section in which various curves were smoothed (*vide supra*, Chapter VII). This may be restated as follows :—Take successive groups of 2, or 3, or 4 . . . . 10 points, beginning again and again at the ordinates for each of the given abscissæ. Find the centres of gravity of each group; that is, erect an ordinate equal to the average of the ordinates of a group at the point half-way between the ends of the abscissæ of the outside ordinates of the group. Draw a line through the points so obtained. It will

---

[*] *Cf.* Edgeworth, in *Statistical Journal*, 1899, p. 381, and the references there given.

be found that this line satisfies all the conditions laid down. An example of this method is given in the diagram facing p. 134.

(3) In another method the original figures are smoothed till the differences of the fourth or fifth or higher orders vanish; and then the ordinary formulæ of interpolation are applied.

Thus in example 1, on p. 233, rewrite the table thus :—

Wages above 15s.	Smoothed Numbers.	Corrected Differences.		
		1st.	2nd.	3rd.
Up to 20s.	296	$303 + a$		
,, 25s.	$599 + a$	$205 + b$	$-98 - a + b$	$7 - 3b$
,, 30s.	$804 + a + b$	$114 - a - b$	$-91 - a - 2b$	$25 + 2a + 3b$
,, 35s.	918	48	$-66 + a + b$	
,, 40s.	966			

If we put $b = 2\frac{1}{3}$, $a = -16$, the third differences vanish, and we have $\triangle_0^1 = 287$, $\triangle_0^2 = -79\frac{2}{3}$, $\triangle_0^3 = \triangle_1^3 = 0$; when $x = 25$, $y = 583$, and when

$$x = 24, \ y = 296 + \tfrac{4}{5} \text{ of } 287 - \tfrac{2}{25} \text{ of } (-79\tfrac{2}{3}) = 531 \cdot 97$$

so that the number earning as much as 24s. and not so much as 25s. is now found to be 51, instead of 52.

The corrections may be applied to any of the original figures.

We need to solve only one more equation to complete our table from 20s. to 30s.

When $x = 23$, $y = 296 + \tfrac{3}{5}$ of $287 + \tfrac{3}{25}$ of $79\frac{2}{3}$. The difference between this and the value of y, when $x = 24$, is $\frac{1}{5}$ of $287 - \tfrac{1}{25}$ of $79\frac{2}{3} = 54 \cdot 21$.

We have therefore the following table, where the figures in italics have already been calculated, while the others are added on the assumption that the third differences are zero.

Wages.	Numbers.	Differences.		
		1st.	2nd.	3rd.
Up to 20s.	*296*		...	...
,, 21s.	360	63·75	3·18	...
,, 22s.	420	60·57	3·18	0
,, 23s.	478	57·39	3·18	0
,, 24s.	532	*54·21*	3·18	0
,, 25s.	*583*	*51·03*	3·18	0
,, 26s.	631	47·85	3·18	0
,, 27s.	676	44·67	3·18	0
,, 28s.	717	41·49	3·18	0
,, 29s.	755	38·31	3·18	0
,, 30s.	790	35·13	...	...

If we had taken the second differences more exactly, we should have obtained $804 + a + b = 790\frac{1}{3}$ for the last figure as in the previous table.

This method of writing down many figures when the significant differences have been found can be very generally applied also in the cases where the data are exact.

(4) Another method, involving higher mathematics, would be discussed more suitably after the section devoted to the law of error; a brief explanation with a useful formula may, however, be offered here.

Suppose we have five consecutive points $(-2, y_{-2})$, $(-1, y_{-1})$, $(0, y)$, $(1, y_1)$, $(2, y_2)$ given.

A parabola of the fourth order could be drawn through these five points, but would have two points of inflexion. A great number of parabolas of the third order can be drawn near all the points, having no points of inflexion, and satisfying all the ordinary conditions of interpolation.

Borrowing a principle from the method of least squares,[*] we assume that if the coefficients of the parabola

$$y = a + bx + cx^2 + dx^3$$

are chosen so as to make the quantity

$$\Sigma(a + bx + cx^2 + dx^3 - y)^2$$

(where the summation extends over the five years of values of x and y) a minimum, the parabola so determined will be the best for the purpose.

For the necessary mathematical analysis, Professor Darwin's paper *On Fallible Measures*,[†] from which this method is taken, should be consulted.

The following equation is obtained—

$a = y_o - \frac{3}{35} \times \triangle_o{}^4$, where $\triangle_o{}^4$ is the difference of the fourth order for the y's.

Now replace the point (o, y) by the intersection of its ordinate with the parabola, that is by (o, a), where a has the value just given, that is, diminish y by the quantity $\frac{3}{35} \cdot \triangle_o{}^4$. Repeat the same process for each point on the original line, taking it as the middle of a group of 5, and a smooth curve lying very near all the original points is obtained.

Thus we may smooth line C in diagram facing p. 146.

---

[*] See Part II, Appendix, Note 10.
[†] See *Phil. Mag. and Journal*, July 1877.

Imported Wheat per head of the Population.	Differences.				Smoothed Figures.
1890 226 (lbs.)					
	18				
1891 244	1	−17			
			19		
1892 245	3	2		−16	245+$\frac{3}{35}$ of 16=246½
			3		
1893 248	8	5		13	248−$\frac{3}{35}$ of 13=247
			16		
1894 256	29	21		−94	256+$\frac{3}{35}$ of 94=264
			−78		
1895 285	−28	−57		134	285−$\frac{3}{35}$ of 134=263½
			56		
1896 257	−29	−1		−16	257+$\frac{3}{35}$ of 16=253½
			40		
1897 228	+10	39			
1898 238					

The statistics of wheat consumption are inexact because of the variation of the stocks at the end of each year, of which no record was available. Hence it is reasonable to regard the numbers as subject to amendment and smooth off irregularities.

(5) A more general problem of interpolation is to find an algebraic formula, other .than the parabolic equation so far used, which expresses a whole series or group. A short introduction to such formula will be found in Part II, Chap. V, below.

Note.—Formula (λ) is due to Professor Everett, who gave the general term and proof (*Quarterly Journal of Pure and Applied Mathematics*, No. 128, 1901, formula G). A proof can be obtained as follows :—

If $f(x) = \cosh\left(2q \sinh^{-1}\frac{x}{2}\right)$, it is readily shown that $f^{n+2}(o) = (q^2 - \frac{1}{4}n^2)f^n(o)$, and thence by Maclaurin's Theorem the expansion of $f(x)$ is—

$$1 + \frac{1}{2!}q^2 x^2 + \frac{1}{4!}q^2(q^2-1)x^4 + \frac{1}{6!}q^2(q^2-1^2)(q^2-2^2)x^6 + \ldots = \cosh\left(2q \sinh^{-1}\frac{x}{2}\right)$$

After differentiating and dividing by qx we obtain—

$$q + \frac{1}{3!}q(q^2-1)x^2 + \frac{1}{5!}q(q^2-1^2)(q^2-2^2)x^4 + \ldots = \frac{2}{x\sqrt{4+x^2}}\sinh\left(2q \sinh^{-1}\frac{x}{2}\right)$$
$$= \frac{\sinh(qhD)}{\sinh(hD)}, \text{ where } x = 2\sinh\left(\frac{hD}{2}\right)$$

In the notation of p. 228, $\delta_o^2 = (e^{hD} - 2 + e^{-hD})y_o = \left(e^{\frac{hD}{2}} - e^{-\frac{hD}{2}}\right)^2 y_o$, so that the operator $\delta = 2\sin h\left(\frac{hD}{2}\right)$.

$$y_{ph} = e^{phD}(y_o) \text{ and } y_1 = e^{hD}(y_o).$$
$$e^{phD} = \text{, identically, } \{e^{(p+1)hD} - e^{(p-1)hD}\} \div (e^{hD} - e^{-hD})$$
$$= \{e^{hD} - e^{-phD} + (e^{hD} - e^{-phD})e^{hD}\} \div (e^{hD} - e^{-hD}),$$

since $p+q=1$,
$$= \frac{\sinh(qhD)}{\sinh(hD)} + \frac{\sinh(phD)}{\sinh(hD)}e^{hD}.$$
$$\therefore y_{ph} = \frac{\sinh(qhD)}{\sinh(hD)}y_o + \frac{\sinh(phD)}{\sinh(hD)}y_1.$$

$\varkappa$ in the above series is identified as δ, and we have, using the series first as expressing operators on y_0, and secondly (after p is written for q) as expressing operators on y_1,

$$y_{p\lambda} = qy_0 + \frac{1}{3!}q(q^2-1)\delta_0^2 + \frac{1}{5!}q(q^2-1)(q^2-2^2)\delta_0^4$$
$$+ \frac{1}{7!}q(q^2-1^2)(q^2-2^2)(q^2-3^2)\delta_0^6 + \ldots$$
$$+ py_1 + \frac{1}{3!}p(p^2-1)\delta_1^2 + \frac{1}{5!}p(p^2-1^2)(p^2-2^2)\delta_1^4$$
$$+ \frac{1}{7!}p(p^2-1^2)(p^2-2^2)(p^2-3^2)\delta_1^6 + \ldots;$$

that is formula (λ) generalized.

For further information on the subject of interpolation, the reader is referred to Dr. Farr's *Life Table* (No. 3), 1864, Boole's *Finite Differences*, *Text-Book of Institute of Actuaries*, Part II., p. 420 *seq.*, Rice's *Theory and Practice of Interpolation*, 1899, Merrifield *On Quadratures and Interpolation* (British Association Report, 1880), Chauvenet's *Spherical and Practical Astronomy* (Chap. II.), Woolhouse in the *Assurance Magazine* (Vols. XI., XII.), Professor J. D. Everett *On the Algebra of Difference Tables* (Quarterly Journal of Mathematics, No. 124, 1900), *On a Central-difference Interpolation Formula* (British Association Report, 1900), and in the Journal of the Institute of Actuaries, January 1901, and Dr. W. F. Sheppard's Papers *On Central Difference Formulæ* (Proceedings of the London Mathematical Society, Vol. XXXI., Nos. 707–710), and *On the Use of Auxiliary Curves in Statistics of Continuous Variation* (Statistical Journal, September 1900). In these other references will be found.

PART II.

APPLICATIONS OF MATHEMATICS TO STATISTICS.

PART II.

APPLICATIONS OF MATHEMATICS TO STATISTICS.

CHAPTER I.

INTRODUCTORY. FREQUENCY CURVES.

Introductory.

MATHEMATICAL processes are essential in very many parts of the statistical field, and in the first part of this book algebraic methods have been used for the generalisation of arithmetical results and for the simpler cases of interpolation. There are, however, many classes of problems which necessitate mathematical treatment of a rather special nature, and it is to the consideration of some of these that this second part is devoted. The whole field is too wide to cover, and selection has been made of those methods which are fundamental and of those problems which are of direct interest to students of political economy and allied sciences. Essentially the same methods are needed for statistical problems in medicine, biology and other sciences, and their use can be followed in the appropriate journals. Here it has seemed best to keep, as a general principle, to those questions which have arisen in connection with economic and social investigation, and to take examples mainly from this limited region.

So far as the manifold and diverse applications can be classified, they fall into three groups: (1) the systematic description of groups, (2) the measurement of relationship between phenomena, (3) the measurement of the precision of results obtained by a process of sampling. The background of the great part of the relevant analysis is the theory of chance,

carried to a point which is reached only by the relatively small number of mathematicians who have specialised in that subject. Since it cannot be assumed that readers are familiar with any but the simpler cases of algebraic probability * 'and there are no familiar text-books in English to which reference can be made, it has been necessary to devote a good deal of space to purely mathematical treatment ; but an effort has been made to render the treatment intelligible to those who have had some mathematical training, but are not specialists in the subject. Thus where possible the proofs have been given without the use of the Infinitesimal Calculus ;' the results have been stated as clearly as possible in words and illustrated by arithmetical examples ; the simplest cases have been dealt with first to elucidate the processes and results, while the more general treatment has been given in outline with reference to papers or journals where a complete analysis has been found. Non-mathematical readers are recommended to omit the parts printed in small type. In the Appendix are collected some theorems whose proofs are not elsewhere very easily accessible, and to it are relegated some parts of the analysis which are too unwieldy for the text.

Frequency Groups and Curves.

The remainder of this chapter is devoted to the systematic measurement of frequency groups.
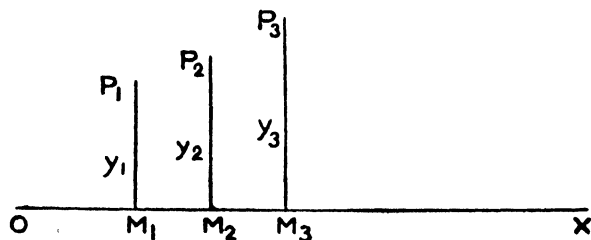
Let there be any group of measurements such that, an axis Ox being taken on which a scale is marked, y_1 instances are found to have the measurement x_1, y_2 the measurement x_2, and so on; then the group can be represented as in the diagram, where $OM_1 = x_1$, $M_1P_1 = y_1$, etc.

---

* For elementary treatment, see Whitworth's *Choice* and *Chance*.

It is not necessary that the grades M_1M_2, M_2M_3 . . . should be equal.

Let n be the whole number in the group, so that

$$n = y_1 + y_2 + \ldots$$

Then the "frequencies" of observations at x_1, x_2, etc., are

$$\frac{y_1}{n}, \frac{y_2}{n}, \text{ etc.}$$

If the points P_1, P_2, P_3 . . . can be regarded as lying on a continuous curve, then their locus is a "frequency curve."

If the measurements do not fall into grades, or in sub-groups at particular values, but each observation has a distinctive measurement, then the group can be represented by a loaded axis on which each item is marked by a dot,

and a great part of the following formulæ is applicable to such a loaded line as well as to a frequency curve.

Measurements of the members of a group are frequently massed in grades (as 20–25, 25–30, . . . years) or originally made to the nearest unit (as 55–56, 56–57 . . . inches). In such cases the number in each grade is approximately represented by a rectangle (as $M_2M_3R_2Q_2$ on the grade M_2M_3).
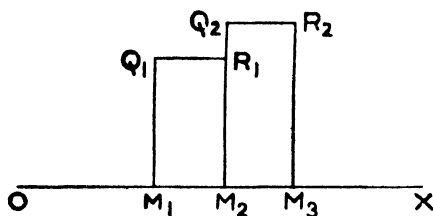
Let h be the breadth of each grade, x_1, x_2 . . . the abscissæ of their middle points, y_1, y_2 . . . the altitudes of the rectangles, and y_1h, y_2h . . . the numbers recorded in the grades. Then

$$n = y_1h + y_2h + \ldots, \quad = y_1 + y_2 + \ldots$$

if h is taken as the unit.

The frequencies in the grades are

$$\frac{y_1h}{n}, \frac{y_2h}{n} \ldots \text{ etc.}$$

If a continuous curve can be defined and constructed so that the parts of its area standing on M_1M_2, M_2M_3 . . . are proportional to y_1h, y_2h . . ., then this is the frequency curve of the group.

Variation is a general law of nature and is found in most human affairs, so that large scale observations usually lead to frequency groups. Four classes can be distinguished : (*a*) where every member of a group has been measured, *e.g.* the wages of every adult male working in a trade ; (*b*) observations of samples selected from a group, *e.g.* the number of children in each of 1,000 families chosen in a town where there are 50,000 families, or the measurement of leaves of a tree of a particular kind ; (*c*) repeated measurements of a physical quantity (*e.g.* of the declination of a star) where the variations are due to instrumental *errors* ; (*d*) the mathematical probabilities of various numbers of successes (*e.g.* the chances of obtaining 1, 2, 3 . . . heads when 50 coins are tossed) or the frequencies of events whose magnitude depends on an unknown complex of causes.

To whichever class the phenomena belong, the same general method of describing the group is appropriate. This method is to select certain algebraic functions of the x's and y's and to evaluate them for the particular group. The group is in fact described (1) by determining a central position, (2) by measuring the dispersion of the observations from this centre, (3) by measuring any want of symmetry about its centre, (4) by further measurements depending on the shape of the diagram which represents the group.

For the central position we can use the arithmetic average, the median, the mode or, in some cases, the geometric mean. The arithmetic average is necessary in most cases in further calculations and must be taken as the usual starting point. The median does not lend itself readily to general algebraic work, is not always known precisely, and need only be calculated for special purposes. The mode is not generally determinable exactly from the observations and the introduction of approximation at the beginning of the calculations should be avoided ; if, however, we have a definite algebraic formula for the group, the mode can be exactly obtained and is often important. (Part I, Chapter V.)

For measurement of dispersion we may use the "probable error," *i.e.*, the half-interquartile range, or the mean deviation,

or the deviation of mean square. Of these the probable error, like the median, can often only be found approximately and is difficult to use systematically in further measurements. The mean deviation,* apart from an ambiguity in the position of the origin from which the deviations are to be measured, .introduces in further work a serious difficulty because the first measurements are taken irrespective of their sign. The deviation of mean square on the other hand is free from all these difficulties, being defined uniquely as the square root of the average of the squares of the deviations of single measurements from their average, and not only is easy to handle algebraically, but also necessarily enters into many calculations. It is called the *standard deviation* and is universally used in mathematical statistics. (Part I, Chapter VI.)

Want of symmetry in a curve is indicated by the want of coincidence of the median, mode and arithmetic average, and by inequality of the distances from the median to the lower and upper quartiles. On any such quantities, which are zero when the group is symmetrical, a measurement can be based ; but the median, mode and quartiles can often only be found approximately, a resulting measurement is specially subject to any imperfections resulting from paucity of observations, and a change in magnitude of an observation has no influence if it does not transfer it across the median or a quartile.

We need a measurement which is sensitive to the position of every observation. It would be possible to take the difference between the mean deviations of observations above and observations below the average, but this would not lead to a formula readily put in line with other systematic measurements. It is found that the deviation of mean cube (the average of the third powers of the deviations of observations from their average, taken positively or negatively as they occur) is free from all difficulties, and it is evidently sensitive to all want of symmetry or " skewness."

In measuring deviation it is natural and usual to express the result in concrete terms as so many inches, lbs., or other units, and the standard deviation, probable error, and mean deviation are so expressed. But in measuring skewness there is no obvious concrete unit and it is convenient to construct the measurement so as to be independent of the unit used ; this is obtained by expressing the deviation of each observa-

tion from the average as a multiple of the standard deviation ; thus if x is a measurement, $\bar{x}$ the average, and σ the standard deviation, the quantities averaged are $\left(\dfrac{x-\bar{x}}{\sigma}\right)^3$, and the resulting measurement of skewness is $\dfrac{1}{n}\left[\text{sum of all values}\right.$ of $\left(\dfrac{x-\bar{x}}{\sigma}\right)^3\right]$. This evidently gives a sensitive measurement, but on no obvious scale, and it is only by experience of the shapes of curves and the resulting measures of skewness that these measures acquire an intelligible meaning.

Further measurements can be obtained from the mean fourth, fifth, and higher powers. These have been generalised by Professor Karl Pearson in his system of *moments*. The 1st, 2nd, 3rd . . . moments are the mean of the first, second, third . . . powers of the deviations ; the deviations may be measured from any point and the resulting moments are with respect to that point ; but the arithmetic average is generally taken as the centre from which measurements are made, and moments with regard to other points are only used to facilitate calculation.

In Part I, Chapter V, it was explained that an average was used as a compact way of describing a group, especially when it was desired to compare or contrast two groups. This conception has now been developed, and we have a systematic way of describing the essential characteristics by three or more symbols, which measure the average, the standard deviation, the skewness and further analogous quantities. As soon as the meanings and scales of these measurements are appreciated, we may dispense with the original data (keeping them only for reference or as diagrams), express groups in a concentrated form, and base calculations showing the relations of groups to each other on these quantities which are specially adapted to mathematical treatment.

The system is not of universal applicability, and in Chapter V are given examples of other methods suitable for particular classes of groups.

Notation of Moments. •

The notation and nomenclature used here are as follows :—

$$m'_t = \frac{1}{n}(x_1^t y_1 + x_2^t y_2 + \ldots) = S(x^t y) \div n. \qquad . \quad . \quad (1)$$

is called the tth *moment* of the group about its origin.

$$n = Sy \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$
$$m_1' = \bar{x} = Sxy \div Sy \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

is the average of the group.

$$m_t = S(x - \bar{x})^t y \div n \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (4)$$

is the tth moment about the average.

Then

$$nm_2 = S(x - \bar{x})^2 y = Sx^2 y - 2\bar{x}Sxy + \bar{x}^2 Sy = nm_2' - 2\bar{x}.n\bar{x} + n\bar{x}^2$$

and

$$\therefore \ m_2 = m_2' - \bar{x}^2 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (5)$$

$$\sigma = \sqrt{m_2} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (6)$$

is called the *standard deviation*, as defined above.

$$nm_3 = Sx^3 y - 3\bar{x}Sx^2 y + 3\bar{x}^2 Sxy - n\bar{x}^3$$
$$m_3 = m_3' - 3\bar{x}m_2' + 2\bar{x}^3 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (7)$$

m_3 is zero in a symmetrical curve. To obtain a convenient measurement of want of symmetry or skewness the abscissæ are expressed as multiples of the standard deviation, thus eliminating the concrete unit of measurement.

Thus $\qquad \kappa^* = S\left\{\left(\dfrac{x - \bar{x}}{\sigma}\right)^3 y\right\} \div n = \dfrac{m_3}{\sigma^3} \quad . \quad . \quad . \quad . \quad (8)$

is a measurement of skewness.

Similarly, $\qquad m_4 = m_4' - 4\bar{x}m_3' + 6\bar{x}^2 m_2' - 3\bar{x}^4 \quad . \quad . \quad . \quad (9)$

is the fourth moment, and

$$\kappa_2 = \frac{m_4}{\sigma^4} = \frac{m_4}{m_2^2} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (10)$$

gives a measurement independent of the unit.

---

* This symbol is introduced in this book in place of letters formerly used to measure skewness. It is believed that it will be found convenient.

The standard deviation being given, the more the members of the group are dispersed from the centre, the greater is κ_2. In the particular case of the normal curve of error (p. 269), $\kappa_2 = 3$. If without altering σ the central height is depressed and the outlying parts pushed further out than in the normal curve, then $\kappa_2 < 3$.

Professor Karl Pearson uses $\beta_1 = \dfrac{m_3{}^2}{m_2{}^3}$, so that $\sqrt{\beta_1} = \kappa$ as given above, and he and Mr. Yule use a more elaborate formula for skewness. Also he writes μ_t instead of m_t, and β_2 for κ_2. Professor Edgeworth, following earlier practice, frequently uses $c = \sqrt{2m_2}$ (called the *modulus*), for the unit of reduction instead of σ, so that $c = \sigma\sqrt{2}$. On the whole the saving of complexity in some formulæ by the use of c may be held not to compensate the use of an additional letter, for in any case the standard deviation must be used.

Edgeworth also uses j for $\dfrac{m_3}{c^3}$, so that $\kappa = 2\sqrt{2}j$, and i for

$$\frac{m_4}{c^4} - \frac{3}{4} = \frac{1}{4}\left(\frac{m_4}{\sigma^4} - 3\right) = \frac{1}{4}(\kappa_2 - 3). \text{ Then } i \text{ is zero in the normal}$$

curve of error.

Illustrations of the Calculation of Moments.

In the following examples methods of calculating the essential measurements $\bar{x}$, σ, κ, κ_2 are given.

In very few cases has it been found necessary or expedient to use higher moments than the fourth for descriptive work, and it is well that this is so, for the errors incident to the obtaining of higher moments from actual observations are generally so considerable as to render them useless.

1. In the first example a fairly homogeneous group of physical measurements is taken, viz., the weights of 3,404 boys of nearly the same age. If their heights (given on p. 385) were symmetrically distributed, it is to be expected that their weights would show a positive skewness, and in fact $\kappa = \cdot643$. One boy of exceptional physique (height 5 ft. 4 in., weight 14 stones) is excluded in the calculation of moments. The curve is not far removed from normality, for $\kappa_2 - 3$ equals only $\cdot457$.

WEIGHT OF BOYS, 14 TO 15 YEARS OF AGE, GRANTED EMPLOYMENT CERTIFICATE IN NEW YORK.

Weight. lbs.	Scale. x	Number. y	xy	x^2y	x^3y	x^4y
65–	− 7	3	− 21	147	− 1,029	7,203
70–	− 6	9	− 54	324	− 1,944	11,664
75–	− 5	142	− 710	3,550	− 17,750	88,750
80–	− 4	301	− 1,204	4,816	− 19,264	77,056
85–	− 3	289	− 867	2,601	− 7,803	23,409
90–	− 2	380	− 760	1,520	− 3,040	6,080
95–	− 1	416	− 416	416	− 416	416
100–	0	404	—	0	—	0
105–	1	315	+ 315	315	+ 315	315
110–	2	320	+ 640	1,280	+ 2,560	5,120
115–	3	262	+ 786	2,358	+ 7,074	21,222
120–	4	221	+ 884	3,536	+ 14,144	56,576
125–	5	131	+ 655	3,275	+ 16,375	81,875
130–	6	76	+ 456	2,736	+ 16,416	98,496
135–	7	52	+ 364	2,548	+ 17,836	124,852
140–	8	20	+ 160	1,280	+ 10,240	81,920
145–	9	29	+ 261	2,349	+ 21,141	190,269
150–	10	14	+ 140	1,400	+ 14,000	140,000
155–	11	10	+ 110	1,210	+ 13,310	146,410
160–	12	2	+ 24	288	+ 3,456	41,472
165–	13	2	+ 26	338	+ 4,394	57,122
170–	14	5	+ 70	980	+ 13,720	192,080
175–	15	1	+ 15	225	+ 3,375	50,625
		3,404	+ 4,906 − 4,032	37,492	+ 158,356 − 51,246	1,502,932
			+ 874		+ 107,110	

The origin is taken at 102·5, and the unit as 5 lbs.

$$m_1' = \bar{x} = \frac{874}{3404} = ·2568 \qquad m_1 = 0. \text{ Average } 102·5 + ·2568 \times 5 = 103·784 \text{ lbs.}$$

$$m_2' = \frac{37492}{3404} = 11·014 \qquad m_2 = m_2' - \bar{x}^2 = 10·948$$

$$m_3' = \frac{107110}{3404} = 31·466 \qquad m_3 = m_3' - 3\bar{x}m_2' + 2\bar{x}^3 = 23·01$$

$$m_4' = \frac{1502932}{3404} = 441·519 \qquad m_4 = m_4' - 4\bar{x}m_3' + 6\bar{x}^2m_2' - 3\bar{x}^4 = 413·542$$

m_2 corrected,[*] $= 10·948 - \frac{1}{12} = 10·865.$ $\sigma = \sqrt{m_2} = 3·296, \text{ i.e. } 16·48 \text{ lbs.}$

m_4 corrected,[*] $= m_4 - \frac{1}{2}m_2 + \frac{7}{240} = 413·542 - 5·474 + ·029 = 408·10$

$\kappa = \frac{m_3}{\sigma^3} = ·643 = \sqrt{\beta_1}, \quad \kappa_2 = \frac{m_4}{m_2^2} = 3·457 = \beta_2$

$c = 4·661 \quad j = \frac{m_3}{c^3} = ·227 \quad i = ·114.$

---

[*] Sheppard's corrections, see Appendix, Note 5, p. 439.

In the above table and in similar calculations it is assumed that the numbers in each grade can be treated as if they were all at the centre of the grade. Unless the grading is very fine, this exaggerates perceptibly the second and fourth moments, while if the numbers in the extreme grades are small the first and third are little affected. If the breadth of the grade is h and not taken as unity, the corrected moments are $m_2 - \frac{1}{12}h^2$ and $m_4 - \frac{1}{2}h^2m_2 + \frac{7}{240}h^4$.

2. Sauerbeck's 45 index numbers measure the movement of prices of separate commodities, while their average measures the general price movement. The 45 numbers may be regarded as measurements of the general movement subject to individual chance deviations, and therefore form a frequency group, whose standard deviation can be used to measure the precision of the average. The group is moderately unsymmetrical. The number of cases is so small that it is not worth while to calculate the 4th moment.

SAUERBECK'S INDEX NUMBERS OF 45 COMMODITIES IN 1916.

Numbers.	x	x^2	x^3	Numbers.	x	x^2	x^3
68	− 68	4,624	− 314,432	138	+ 2	4	8
71	− 65	4,225	− 274,625	148	+ 12	144	1,728
84	− 52	2,704	− 140,608	148	+ 12	144	1,728
86	− 50	2,500	− 125,000	153	+ 17	289	4,913
93	− 43	1,849	− 79,507	154	+ 18	324	5,832
96	− 40	1,600	− 64,000	154	+ 18	324	5,832
100	− 36	1,296	− 46,656	157	+ 21	441	9,261
100	− 36	1,296	− 46,656	159	+ 23	529	12,167
101	− 35	1,225	− 42,875	159	+ 23	529	12,167
104	− 32	1,024	− 32,768	160	+ 24	576	13,824
104	− 32	1,024	− 32,768	161	+ 25	625	15,625
107	− 29	841	− 24,389	163	+ 27	729	19,683
114	− 22	484	− 10,648	163	+ 27	729	19,683
114	− 22	484	− 10,648	166	+ 30	900	27,000
119	− 17	289	− 4,913	168	+ 32	1,024	32,768
121	− 15	225	− 3,375	169	+ 33	1,089	35,937
125	− 11	121	− 1,331	172	+ 36	1,296	46,656
128	− 8	64	− 512	173	+ 37	1,369	50,653
128	− 8	64	− 512	174	+ 38	1,444	54,872
131	− 5	25	− 125	183	+ 47	2,209	103,823
132	− 4	16	− 64	197	+ 61	3,721	226,981
135	− 1	1	− 1	202	+ 66	4,356	287,496
135	− 1	1	− 1				
				22	629	22,795	988,637
23	− 632	25,982	− 1,256,414	23	− 632	25,982	− 1,256,414
				45	− 3	48,777	− 267,777

Origin at 136.

$$\bar{x} = -\frac{3}{45}. \text{ Average } 136 - \frac{3}{45} = 135 \cdot 93$$

$$m_2' = \frac{48777}{45} = 1083 \cdot 933 \qquad m_2 = m_2' - \bar{x}^2 = 1083 \cdot 929. \quad \sigma = \sqrt{m_2} = 32 \cdot 9$$

$$m_3' = -\frac{267777}{45} = -5951 \qquad m_3 = m_3' - 3\bar{x}m_2' + 2\bar{x}^3 = -5734$$

$$\beta = \frac{m_3}{\sigma^3} = -\cdot 161$$

3. OBSERVATIONS OF THE RIGHT ASCENSION OF THE POLE STAR.*

Seconds from assumed mean	x	Number of Observations. y	xy	x^2y	x^3y	x^4y
+3·0	6	1	6	36	216	1,296
+2·5	5	5	25	125	625	3,125
+2·0	4	16	64	256	1,024	4,096
+1·5	3	38	114	342	1,026	3,078
+1·0	2	63	126	252	504	1,008
+0·5	1	72	72	72	72	72
0·0	0	82	—	0	—	0
-0·5	-1	73	- 73	73	- 73	73
-1·0	-2	61	-122	244	- 488	976
-1·5	-3	36	-108	324	- 972	2,916
-2·0	-4	21	- 84	336	-1,344	5,376
-2·5	-5	12	- 60	300	-1,500	7,500
-3·0	-6	6	- 36	216	-1,296	7,776
-3·5	-7	1	- 7	49	- 343	2,401
		487	+407 / -490	2,625	3,467 / -6,016	39,693
			- 83		-2,549	

$$= - \cdot 170$$
$$m_2 = 5 \cdot 390 - \cdot 029 = 5 \cdot 361. \qquad \sigma = 2 \cdot 3$$
$$m_3 = - 5 \cdot 234 - 3(- \cdot 170) \times 5 \cdot 390 + 2(- \cdot 170)^3 = - 2 \cdot 49 \qquad \kappa = - \cdot 2$$
$$m_4 = 81 \cdot 505 - 4(- \cdot 170)(-5 \cdot 234) + 6(\cdot 170)^2(5 \cdot 390) - 3(\cdot 170)^4 = 78 \cdot 88 \qquad \kappa_2 = 2 \cdot 7$$

These observations have been frequently used in discussing how far physical observations can be expressed by the normal curve. The results are nearly symmetrical, but since $\kappa_2 < 3$ there is an under-concentration near the average.

4. The following example shows how a table of chances can be treated as a frequency group ; an unsymmetrical case has been selected, namely the chance of obtaining sixes in a throw of 12 dice ; *e.g.*, the chance of exactly 3 sixes is

$$_{12}C_3 \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^9, \quad \text{see p. 262.}$$

Number of Sixes. x	Chance in 12 throws. y
0	244,140,625 ÷ 6^12
1	585,937,500 ,,
2	644,531,250 ,,
3	429,687,500 ,,
4	193,359,375 ,,
5	61,875,000 ,,
6	14,437,500 ,,
7	2,475,000 ,,
8	309,375 ,,
9	27,500 ,,
10	1,650 ,,
11	60 ,,
12	1 ,,
	2,176,782,336 ,,

$\bar{x} = 2$
$m_2 = 1\frac{2}{3}$
$m_3 = 1\frac{1}{3}$
$m_4 = 8\frac{11}{18}$
$\sigma = 1 \cdot 29$
$\kappa = \cdot 516$
$\kappa_2 = 3 \cdot 1$

* Quetelet, *Lettres sur la théorie des probabilités*, p. 128.

5. If digits are selected at random their average may be expected to tend to 4·5. The group in the table below shows the result of selecting 400 groups of 25 each from the last digits in 7 figure logarithm tables. The group is somewhat unsymmetrical and $\kappa_2 > 3$.

<div align="center">SUM OF 25 DIGITS, DIVIDED BY 5.</div>

Difference from 22·5.				Number of Cases.	
Over 9	.	.	.	1	
8 to 9	.	.	.	5	
7 „ 8	.	.	.	9	
6 „ 7	.	.	.	5	
5 „ 6	.	.	.	12	
4 „ 5	.	.	.	10	
3 „ 4	.	.	.	15	With origin 23,
2 „ 3	.	.	.	36	$\bar{x} = -·2575$; average, 22·7425
1 „ 2	.	.	.	48	$m_2 = 8·8662$; corrected, 8·783
0 „ 1	.	.	.	57	$\sigma = 2·964$
0 „ −1	.	.	.	62	$m_3 = 13·584$; $\kappa = ·522$
−1 „ −2	.	.	.	58	$m_4 = 274·24$; corrected, 269·8
−2 „ −3	.	.	.	39	$\kappa_2 = 3·50$
−3 „ −4	.	.	.	17	
−4 „ −5	.	.	.	13	
−5 „ −6	.	.	.	10	
−6 „ −7	.	.	.	2	
−7 „ −8	.	.	.	1	
				400	

Mr. Elderton [*] gives a method of calculating moments specially suited for work on an adding and multiplying machine, which may be expressed as follows in the notation of this chapter.

Let $y_1, y_2 \ldots y_t$ be the frequencies at $x = 1, 2 \ldots t$.

Write $_0S_1 = y_t,\ _0S_2 = y_t + y_{t-1},\ \ldots _0S_t = y_t + y_{t-1} + \ldots + y_1.$

Also write

$_1S_2 = {_0S_1} + {_0S_2},\ _1S_3 = {_0S_1} + {_0S_2} + {_0S_3},\ \ldots, {_1S_t} = {_0S_1} + {_0S_2} + \ldots {_0S_t},$
and $_2S_2 = {_1S_1} + {_1S_2}, \ldots, {_2S_t} = {_1S_1} + {_1S_2} + \ldots + {_1S_t},$ and so on.

$_0S_t =$ number of observations $= n$

$_1S_t = ty_t + (t-1)y_{t-1} + \ldots + 1 . y_1 = n\bar{x},$

where $\bar{x}$ is the average, $= nm_1'$.

$$_2S_t = (1+2+\ldots+t)y_t + (1+2+\ldots+\overline{t-1})y_{t-1}+\ldots+(1+2)y_2+y_1$$

$$= \frac{t(t+1)}{2}y_t + \frac{(t-1)t}{2}y_{t-1} + \ldots + \frac{1\cdot 2}{2}y_1 = \frac{n}{2}(m_2' + m_1'),$$

where m_1', m_2' ... are moments about the origin.

$$_3S_t = \frac{1}{2}\{1\cdot 2 + 2\cdot 3 + \ldots t(t+1)\}y_t + \frac{1}{2}\{1\cdot 2 + \ldots (t-1\ t\}y_{t-1}+\ldots$$

$$= \frac{1}{6}\{t(t+1)(t+2)y_t + (t-1)t(t+1)y_{t-1}+\ldots+1\cdot 2\cdot 3y_1\}$$

$$= \frac{n}{6}(m_3' + 3m_2' + 2m_1'), \text{ and}$$

$$_4S_t = \frac{n}{24}(m_4' + 6m_3' + 11m_2' + 6m_1').$$

Then by the use of equations 5, 7, and 9 we find

$$m_2 = \frac{2}{n}\cdot {}_2S_t - \bar{x}(1+\bar{x})$$

$$m_3 = \frac{6}{n}\cdot {}_3S_t - 3m_2(1+\bar{x}) - \bar{x}(1+\bar{x})(2+\bar{x})$$

$$m_4 = \frac{24}{n}\cdot {}_4S_t - 2m_3(3+2\bar{x}) - m_2(11+18\bar{x}+6\bar{x}^2) - \bar{x}(1+\bar{x})(2+\bar{x})(3+\bar{x}).$$

The quantities $_1S_t$, $_2S_t$, $_3S_t$, $_4S_t$ are quickly obtained by repeated addition. The process is exhibited sufficiently by working out the moments of Example 5 (p. 256) by this method.

x is measured from the origin 14; y_x is the number of cases at x. Each term in the column $_0S_x$ is obtained by adding the terms in the previous column that stand to the left and above it; the column $_1S_x$ is obtained similarly from the column $_0S_x$ and so on. $t = 18$. Write $19 - x = x'$.

Sum of digits ÷ 5	x	$y_{x'}$	$_0S_{x'}$	$_1S_{x'}$	$_2S_{x'}$	$_3S_{x'}$
Over 31·5	18	1	1	1	1	1
30·5	17	5	6	7	8	9
29·5	16	9	15	22	30	39
28·5	15	5	20	42	72	111
27·5	14	12	32	74	146	257
26·5	13	10	42	116	262	519
25·5	12	15	57	173	435	954
24·5	11	36	93	266	701	1,655
23·5	10	48	141	407	1,108	2,763
22·5	9	57	198	605	1,713	4,476
21·5	8	62	260	865	2,578	7,054
20·5	7	58	318	1,183	3,761	10,815
19·5	6	39	357	1,540	5,301	16,116
18·5	5	17	374	1,914	7,215	23,331
17·5	4	13	387	2,301	9,516	32,847
16·5	3	10	397	2,698	12,214	45,061
15·5	2	2	399	3,097	15,311	60,372
14·5	1	1	400	3,497	18,808	79,180
Totals		400 $= {}_0S_{18}$	3,497 $= {}_1S_{18}$	18,808 $= {}_2S_{18}$	79,180 $= {}_3S_{18}$	285,560 $= {}_4S_{18}$

$$\bar{F} = \frac{3497}{400} = 8{\cdot}7425 \quad \text{Average} = 22{\cdot}7425$$

$$m_2 = \frac{2}{400} \times 18808 - 8{\cdot}7425 \times 9{\cdot}7425 = 8{\cdot}8662$$

$$m_3 = \frac{6}{400} \times 79180 - 3 \times 8{\cdot}8662 \times 9{\cdot}7425 - 8{\cdot}7425 \times 9{\cdot}7425 \times 10{\cdot}7425 = 13{\cdot}584$$

$$m_4 = \frac{24}{400} \times 285560 - 2 \times 13{\cdot}584 \times 20{\cdot}485 - 8{\cdot}8662 \times 626{\cdot}95 - 10744{\cdot}15 = 274{\cdot}24$$

CHAPTER II.

ALGEBRAIC PROBABILITY AND THE NORMAL CURVE OF ERROR.

Elementary Principles.

THE method and fundamental theorems of algebraic probability may be summarised as follows :—

Suppose that there are N alternative events, any one of which is just as likely to take place as any other, and that one of them is known to have taken place, but we are in complete ignorance which ; further, of the N events suppose that M have a special characteristic and the remaining (N − M) have not ; then the chance that the event that has happened has this characteristic is defined as $\frac{M}{N}$.

Thus, if one card has been drawn from an ordinary pack of 52, the chance that it is a heart is $\frac{13}{52} = \frac{1}{4}$. Here each of the 52 events is so far as we know equally likely, and the skill of the card manufacturer is directed to make the cards of equal weight and with equal friction. We cannot point to any circumstance which tends to give one card rather than another, unless the surface friction of an ace is less than that of a king. In an ideal system there is nothing to distinguish the circumstances that lead to one of the N events rather than another. In the apparatus of fair games of chance this equality is definitely aimed at, and consequently such games supply illustrations of algebraic probability.

Let $p = \frac{M}{N}$; $q = 1 - p = \frac{N - M}{N}$. q is the chance that the characteristic will not be found. If we call the appearance of the characteristic a "success," p is the chance of success, q is the chance of failure ; the odds in favour are p to q, against q to p.

Multiplication of Chances.

If p_1, p_2 are the chances of success in two independent experiments, then $p_1 \times p_2$ can be shown as follows to be the chance of a double success.

In one experiment let there be n_1 equally likely alternative events, and in the other n_2. Write $p_1 = \dfrac{m_1}{n_1}$, $p_2 = \dfrac{m_2}{n_2}$.

By independence here we mean that the result of the first experiment has no effect on the second experiment, so that each of the $n_1 \times n_2$ possible double events is equally likely.

Of these $n_1 \times n_2$ events $m_1 \times m_2$ give a double success

$\qquad\qquad\qquad\qquad m_1 \times (n_2 - m_2)$ give success and failure

$\qquad\qquad\qquad\qquad (n_1 - m_1) \times m_2$ give failure and success

$\qquad\qquad\qquad\qquad (n_1 - m_1) \times (n_2 - m_2)$ give double failure.

Of $n_1 n_2$ equally likely events $m_1 m_2$ give a double success and the remainder do not. Hence p the chance of double success

$$= \frac{m_1 m_2}{n_1 n_2} = p_1 \times p_2.$$

E.g. the chance that two sixes will be thrown by a pair of dice is $\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$.

If, however, the experiments are not independent, but the result of the first affects the chances in the second, the formula must be modified in the way illustrated by the following example.

If a card is drawn from each of two packs the chance of drawing two aces is $\frac{4}{52} \times \frac{4}{52}$, where $p_1 = \frac{4}{52} = p_2$.

But if the second card is drawn from a pack from which the first has already been taken, we have the following alternatives :—

There are 52×51 possible events.

If an ace is drawn first, there are 3 aces in the remaining 51.

4×3 ways give a double success. 4×48 give success and failure ; 48×4 give failure and success, and 48×47 give double failure.

The chance of a double success is therefore $\frac{4}{52} \times \frac{3}{51} = \frac{1}{221}$.

This problem may also be worked out as follows. There are $_{52}C_2 = \dfrac{52 \cdot 51}{1 \cdot 2}$ pairs in the pack. Of these $_4C_2 = \dfrac{4 \cdot 3}{1 \cdot 2}$ are

two aces. Any pair is as likely to be drawn as any other. Hence the chance of drawing two aces, whether together or consecutively, is $\dfrac{_4C_2}{_{52}C_2} = \dfrac{4 \cdot 3}{52 \cdot 51}$.

The chance of obtaining 8 hearts and 5 cards of other suits in a hand of 13 cards dealt from 52 is

$$\dfrac{_{13}C_8 \times _{39}C_5}{_{52}C_{13}} = \dfrac{13 \cdot 12 \cdot 11 \cdot 10 \cdot 9 \cdot 8 \cdot 7 \cdot 6 \cdot 39 \cdot 38 \cdot 37 \cdot 36 \cdot 35\,(13!)}{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48 \cdot 47 \cdot 46 \cdot 45 \cdot 44 \cdot 43 \cdot 42 \cdot 41 \cdot 40\,(8!)(5!)}$$

$$= \dfrac{105{,}857{,}037}{90{,}716{,}222{,}800} = \dfrac{1}{857} \text{ approx.} = p\,;$$

for there are $_{52}C_{13}$ equally likely hands $= N$; there are $_{13}C_8$ equally likely groups of 8 hearts and $_{39}C_5$ equally likely groups of 5 from other suits, and $\therefore M = _{13}C_8 \times _{39}C_5$, where $p = \dfrac{M}{N}$.

Addition of Chances.

The total 9 can be obtained from the throw of two dice from either of the pairs (3, 6) (4, 5) (5, 4) (6, 3); that is of 36 equally probable events 4 give the result, and the chance is therefore $\frac{4}{36} = \frac{1}{9}$.

This result may also be obtained thus: the chance of throwing 3 is $\frac{1}{6}$, of throwing 6 is $\frac{1}{6}$, and therefore the chance of throwing 3 and 6 is $\frac{1}{36}$. Similarly the chance of throwing (4, 5) (5, 4) and (6, 3) is $\frac{1}{36}$ in each case. The whole chance is the sum of the chances of these alternative double events.

Generally if a success can be obtained *either* from an occurrence whose chance is p_1 followed by one whose chance is p_1', or from successive occurrences whose chances are p_2, p_2' . . . , then the whole chance of a success is

$$P = p_1 p_1' + p_2 p_2' + \ldots .$$

Deduction of the Normal Law of Error.

We can now proceed to a general theorem of great importance alike in the theory of probability itself and in its application to statistics.

Suppose an experiment (*e.g.* throwing dice, drawing a card, or choosing a number) to be such that the chance of success is always p and of failure q, so that $p + q = 1$.

Let the experiment be repeated n times, and consider the chance of obtaining r successes and $\overline{n-r}$ failures. The chance

in an order assigned thus—the first r experiments successes and the rest failures, is

$$p \times p \times \text{to } r \text{ factors} \times q \times q \times \text{to } \overline{n-r} \text{ factors} = p^r \times q^{n-r};$$

and the chance in any other assigned order is the same. The order may be assigned by choosing any r positions for successes in a series of n experiments, *i.e.* in $_nC_r$ ways. Hence the whole chance is $_nC_r \cdot p^r q^{n-r}$.

The chances of $0, 1, 2 \ldots n$ successes are therefore the successive terms of the binomial expansion

$$1 = (q+p)^n = q^n + n \cdot q^{n-1}p + \ldots + _nC_r \cdot q^{n-r}p^r + \ldots + nqp^{n-1} + p^n$$

For example, if $p = \frac{2}{5}$, $q = \frac{3}{5}$ and $n = 10$ we have

r	$_nC_r$	$p^r q^{n-r}$	$_nC_r \cdot p^r q^{n-r}$
0	1	$3^{10} \div 5^{10}$	·006,046,617,6
1	10	2×3^9 ,,	·040,310,784,0
2	45	$2^2 \times 3^8$,,	·120,932,352,0
3	120	$2^3 \times 3^7$,,	·214,990,848,0
4	210	$2^4 \times 3^6$,,	·250,822,656,0
5	252	$2^5 \times 3^5$,,	·200,658,124,8
6	210	$2^6 \times 3^4$,,	·111,476,736,0
7	120	$2^7 \times 3^3$,,	·042,467,328,0
8	45	$2^8 \times 3^2$,,	·010,616,832,0
9	10	$2^9 \times 3^1$,,	·001,572,864,0
10	1	2^{10} ,,	·000,104,857,6

$$1 \cdot 000,000,000,0$$
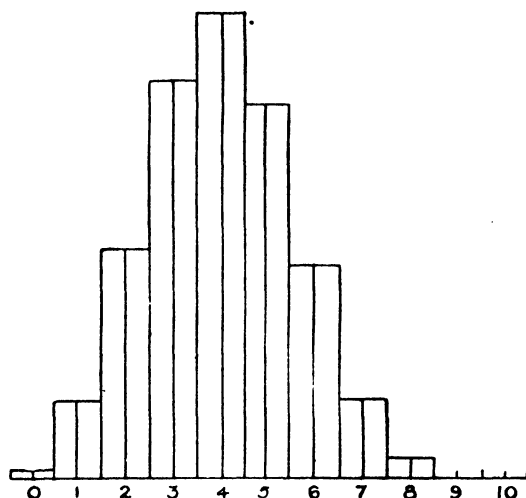
The Vertical scale is expanded 100 fold so that the area of the figure is 100 squares on unit base.

The diagram illustrates the relative chances of different numbers of successes, and exhibits them as a frequency group.

We will first find the moments of the group for general values of p and n. Take the horizontal scale on the diagram as the scale for x.

Suppose the n-fold experiment repeated N times, where N is a very large number. Then the number of times r successes are obtained tends to be $N \times {}_nC_r . q^{n-r}p^r = y_r$, say,

and $y_0 + y_1 + \ldots + y_n = N(q+p)^n = N$, since $p + q = 1$,

$\bar{x} = m_1'$, the first moment about the origin,

$$= (y_0 \times 0 + y_1 \times 1 + \ldots + y_r \times r + \ldots + y_n \times n) \div N$$

$$= n . q^{n-1}p + n(n-1)/2 . q^{n-2}p^2 \times 2 + \ldots + {}_nC_r q^{n-r}p^r \times r + \ldots + p^n \times n$$

$$= np(q+p)^{n-1} = np \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (11)$$

$$m_2' = (y_0 \times 0^2 + y_1 \times 1^2 + \ldots + y_r \times r^2 + \ldots + y_n \times n^2) \div N$$

$$= \sum_0^n r^2 . {}_nC_r . q^{n-r}p^r = \sum\{r(r-1) + r\} \frac{(n)_r}{r!} q^{n-r}p^r$$

$$= n(n-1)p^2 \sum \frac{(n-2)_{r-2}}{(r-2)!} q^{n-r}p^{r-2} + np \sum \frac{(n-1)_{r-1}}{(r-1)!} q^{n-r}p^{r-1}$$

$$= n(n-1)p^2(q+p)^{n-2} + np(q+p)^{n-1} = n(n-1)p^2 + np$$

$$= n^2p^2 + np(1-p) = \bar{x}^2 + npq \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (12)$$

and m_2, the second moment about the average,

$$= m_2' - \bar{x}^2 = npq = np(1-p). \quad . \quad . \quad . \quad . \quad (13)$$

In a similar way

$$m_3' = \sum_0^n r^3 . {}_nC_r . q^{n-r}p^r = n(n-1)(n-2)p^3 + 3n(n-1)p^2 + np . \quad (14)$$

and m_3, the third moment about the average,

$$= m_3' - 3m_2'\bar{x} + 2\bar{x}^3$$

$$= n(n-1)(n-2)p^3 + 3n(n-1)p^2 + np - 3n^3p^3 - 3n^2p^2(1-p) + 2n^3p^3$$

$$= np(2p^2 - 3p + 1) = np(1-p)(1-2p) = npq(q-p) \quad . \quad . \quad (15)$$

$$m_4' = \sum_0^n r^4 . {}_nC_r . q^{n-r}p^r, \text{ and } m_4 \text{ can be shown to equal}$$

$$3(pqn)^2 + pqn(1 - 6pq).$$

Hence, using the formulæ of pp. 251–2, $\sigma = \sqrt{pqn}$, $\beta_1 = \dfrac{(q-p)^2}{pqn}$,

$\kappa_2 = \beta_2 = 3 + \dfrac{1 - 6pq}{pqn}$, $c = \sqrt{2pqn}$, $\kappa = \dfrac{q - p}{\sqrt{pqn}}$, $i = \dfrac{1 - 6pq}{4pqn}$.

The standard deviation varies as $\sqrt{n}$. κ and $\sqrt{\beta_1}$, measurements of skewness, are small when $\sqrt{n}$ is great. $(\kappa_2 - 3)$ and i are small when n is great.

Next consider the chance of r successes and the shape assumed by the diagram when n is increased.

CASE I., when $p = q = \frac{1}{2}$ and n is even $= 2n'$.

Let P_x be the chance of $n' + x$ successes, and therefore $n' - x$ failures.

$$P_x = {}_{2n'}C_{n'+x} \cdot \frac{1}{2^{n'+x}} \cdot \frac{1}{2^{n'-x}} = \frac{(2n')!}{(n'+x)!\,(n'-x)!} \cdot \frac{1}{2^{2n'}}$$

$$= \frac{(2n')!}{n'!\,n'!} \cdot \frac{1}{2^{2n'}} \cdot \frac{n'\,(n'-1)\ldots(n'-x+1)}{(n'+1)\,(n'+2)\ldots(n'+x)}$$

$$P_0 = \frac{(2n')!}{2^{2n'}\cdot n'!\,n'!} = \frac{1}{\sqrt{\pi n'}}, \text{ by Wallis's Theorem, correct}$$

to $\dfrac{1}{n'}$, (Appendix, Note 1 (132)).

$$\therefore\ P_x = \frac{1}{\sqrt{\pi n'}} \cdot \frac{\left(1 - \dfrac{1}{n'}\right)\left(1 - \dfrac{2}{n'}\right)\ldots\left(1 - \dfrac{x-1}{n'}\right)}{\left(1 + \dfrac{1}{n'}\right)\left(1 + \dfrac{2}{n'}\right)\ldots\left(1 + \dfrac{x}{n'}\right)}$$

$$\log\left(P_x\sqrt{\pi n'}\right) = \log\left(1 - \frac{1}{n'}\right) - \log\left(1 + \frac{1}{n'}\right) + \log\left(1 - \frac{2}{n'}\right) - \log\left(1 + \frac{2}{n'}\right) + \ldots$$

$$+ \log\left(1 - \frac{x}{n'}\right) - \log\left(1 + \frac{x}{n'}\right) - \log\left(1 - \frac{x}{n'}\right)$$

$$= -2\left(\frac{1}{n'} + \frac{1}{3n'^3} + \ldots\right) - 2\left(\frac{2}{n'} + \frac{2^3}{3n'^3} + \ldots\right)\ldots$$

$$- 2\left(\frac{x}{n'} + \frac{x^3}{3n'^3} + \ldots\right) - \log\left(1 - \frac{x}{n'}\right).$$

$$= -2\,\frac{1 + 2 + \ldots + x}{n'} - \frac{2}{3} \cdot \frac{1^3 + 2^3 + \ldots + x^3}{n'^3} \ldots$$

$$- \frac{2}{2t+1} \cdot \frac{1^{2t+1} + 2^{2t+1} + \ldots + x^{2t+1}}{n'^{2t+1}} \ldots \therefore -\log\left(1 - \frac{x}{n'}\right),$$

where t is any integer,

$$= -\frac{x(x+1)}{n'} - \frac{2}{3} \cdot \frac{x^2(x+1)^2}{4n'^3} - \dots$$

$$- \frac{2}{2t+1} \cdot \frac{x^{2t+2} + \dots + \dots}{(2t+2)n'^{2t+1}} * - \dots + \left(\frac{x}{n'} + \frac{x^2}{2n'^2} + \dots\right)$$

Write $x = \tau\sqrt{n'} = \tau c$, since from p. 252

$$c^2 = 2pqn = 2 \cdot \tfrac{1}{2} \cdot \tfrac{1}{2} \cdot 2n' = n'.$$

$$\log\left(P_x \cdot c\sqrt{\pi}\right) = -\tau^2 - \frac{\tau}{\sqrt{n'}} - \frac{2}{3n'}\frac{\tau^2}{4}\left(\tau + \frac{1}{\sqrt{n'}}\right)^2 - \dots$$

$$- \frac{2}{(2t+1)(2t+2)}\left(\frac{\tau^{2t+2}}{n'^t} + \dots\right) - \dots + \left(\frac{\tau}{\sqrt{n'}} + \frac{\tau^2}{2n'} + \dots\right)$$

$$= -\tau^2 + \text{terms involving } \frac{1}{\sqrt{n'}}.$$

Hence if $\dfrac{1}{\sqrt{n'}}$ is neglected, as in the value of P_0 above,

$$P_x = \frac{1}{c\sqrt{\pi}}e^{-\tau^2} = \frac{1}{c\sqrt{\pi}}e^{-\frac{x^2}{c^2}} = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}} \quad \dots \quad (16)$$

since σ, the standard deviation, $= c/\sqrt{2}$.

And since $c^2 = \dfrac{n}{2}$, $P_x = \dfrac{\sqrt{2}}{\sqrt{\pi n}}e^{-\frac{2x^2}{n}}$.

CASE II., when p and q are unequal.

Let P_x be the chance of $pn + x$ successes,[†] and therefore $qn - x$ failures.

$$P_x = \frac{n!}{(pn+x)!\,(qn-x)!} \cdot p^{pn+x}q^{qn-x}$$

$$= \frac{n!}{(pn)!\,(qn)!}p^{pn}q^{qn} \frac{qn(qn-1)\dots(qn-x+1)}{(pn+1)(pn+2)\dots(pn+x)} \cdot \frac{p^x}{q^x}$$

$$= P_0 \cdot \frac{\left(1 - \dfrac{1}{qn}\right)\left(1 - \dfrac{2}{qn}\right)\dots\left(1 - \dfrac{x-1}{qn}\right)}{\left(1 + \dfrac{1}{pn}\right)\left(1 + \dfrac{2}{pn}\right)\dots\left(1 + \dfrac{x}{pn}\right)}$$

---

* Appendix 2, formula (133).

† It is assumed for simplicity in the sequel that pn is integral and therefore P_0 the greatest term; since n is large and powers of $\dfrac{1}{n}$ are finally neglected, the proof is not affected.

$$\log(\mathrm{R}_x/\mathrm{P}_0) = \sum_{s=1}^{s=x}\log\left(1 - \frac{s}{qn}\right) - \sum_{s=1}^{s=x}\log\left(1 + \frac{s}{pn}\right) - \log\left(1 - \frac{x}{qn}\right)$$

$$= -\sum_1^x\left(\frac{s}{qn} + \frac{s}{pn}\right) - \sum_1^x\frac{1}{2}\left(\frac{s^2}{q^2n^2} - \frac{s^2}{p^2n^2}\right) - \ldots$$

$$- \sum_1^x\frac{1}{t}\left(\frac{s^t}{q^tn^t} \pm \frac{s^t}{p^tn^t}\right) - \ldots - \log\left(1 - \frac{x}{qn}\right)$$

$$= -\frac{x(x+1)}{2}\cdot\frac{p+q}{pqn} - \frac{x(x+1)(2x+1)}{6}\cdot\frac{p^2 - q^2}{2p^2q^2n^2}$$

$$- \frac{x^2(x+1)^2}{4}\cdot\frac{p^3 + q^3}{3p^3q^3n^3} - \ldots$$

$$- \frac{1}{t}\cdot\frac{x^{t+1} + \ldots}{t+1}\cdot\frac{p^t \pm q^t}{p^tq^tn^t} - \ldots + \left(\frac{x}{qn} + \frac{x^2}{2q^2n^2} + \ldots\right)$$

Write $x = \tau c$, where $c^2 = 2pqn = 2\sigma^2$

$$\log(\mathrm{P}_x/\mathrm{P}_0) = -\frac{\tau^2c^2 + \tau c}{c^2} + \frac{2\tau^3c^3 + 3\tau^2c^2 + \tau c}{3c^4}(q - p)$$

$$- \frac{\tau^4c^4 + 2\tau^3c^3 + \tau^2c^2}{\frac{3}{2}\cdot c^4}\cdot(1 - 3pq) - \ldots$$

$$- \frac{\tau^{t+1}c^{t+1} + \ldots}{t(t+1)2^{-t}c^{2t}}(p^t \pm q^t) - \ldots + \frac{2\tau cp}{c^2} + \frac{2\tau^2c^2p^2}{c^4} + \ldots,$$

$$\text{since } p + q = 1,$$

$$= -\tau^2 + \frac{\tau}{c}\left\{-1 + \frac{2\tau^2}{3}(q - p) + 2p\right\}$$

$$+ \frac{\tau^2}{c^2}\left\{(q - p) - \frac{2\tau^3}{3}(1 - 3pq) + 2p^2\right\}$$

$$+ \text{ terms involving } \frac{1}{c^3}$$

Regard τ as finite; that is, consider only those values of x which are comparable with $\sqrt{pqn}$.

If we neglect $\frac{1}{c}$, $\left(\text{that is, if we neglect } \frac{1}{\sqrt{n}}\right)$, we have

$$\mathrm{P}_x = \mathrm{P}_0 e^{-\tau^2} = \mathrm{P}_0 e^{-\frac{x^2}{c^2}} \quad . \quad . \quad . \quad . \quad . \quad (17)$$

If we keep $\frac{1}{c}$, neglecting $\frac{1}{c^2}$ (that is, neglecting $\frac{1}{n}$), we have

$$\mathrm{P}_x = \mathrm{P}_0 e^{-\tau^2} \cdot e^{-\frac{\tau}{c}(q-p)(1-\frac{1}{3}\tau^2)}$$

$$= \mathrm{P}_0 e^{-\tau^2}\{1 - \frac{q-p}{c}(\tau - \tfrac{2}{3}\tau^3)\}, \text{ since } \frac{1}{c^2} \text{ is neglected,}$$

$$= \mathrm{P}_0 e^{-\frac{x^2}{c^2}}\left\{1 - \frac{q-p}{c}\left(\frac{x}{c} - \frac{2}{3}\cdot\frac{x^3}{c^3}\right)\right\}$$

$$= \mathrm{P}_0 e^{-\frac{x^2}{2\sigma^2}}\left\{1 - \frac{\kappa}{2}\left(\frac{x}{\sigma} - \frac{x^3}{3\sigma^3}\right)\right\}, \quad \ldots \ldots \quad (18)$$

since $c = \sqrt{2} \cdot \sigma = \sqrt{2pqn}$, and $\kappa = \dfrac{q-p}{\sigma}$.

The value of P_0 may be obtained from Stirling's theorem for factorials (Appendix, Note 3 (134)), viz.: $m\,! = m^m\sqrt{2\pi m}\cdot e^{-m+\frac{1}{12m}}$, when $\frac{1}{m^2}$ is neglected, and $= m^m\sqrt{2\pi m}\,e^{-m}$, when $\frac{1}{m}$ is neglected.

$$\mathrm{P}_0 = \frac{n\,!}{(pn)\,!\,(qn)\,!}p^{pn}q^{qn}$$

$$= \frac{n^n}{(pn)^{pn}(qn)^{qn}}\cdot\sqrt{\frac{2\pi n}{2\pi pn \cdot 2\pi qn}}\cdot e^{-n+pn+qn}\cdot p^{pn}q^{qn},$$

$$\text{neglecting } \frac{1}{pn}\cdot\frac{1}{n}, \text{ \&c.,}$$

$$= \frac{1}{\sqrt{2\pi pqn}}, \text{ since } p + q = 1, = \frac{1}{c\sqrt{\pi}} = \frac{1}{\sigma\sqrt{2\pi}}.$$

Now write y for P_x, and we obtain the equations

$$y = \frac{1}{\sqrt{2\pi pqn}}\ e^{-\frac{x^2}{2pqn}} = \frac{1}{c\sqrt{\pi}}\ e^{-\frac{x^2}{c^2}} = \frac{1}{\sigma\sqrt{2\pi}}\ e^{-\frac{x^2}{2\sigma^2}} \quad . \quad . \quad (19)$$

when $\frac{1}{\sqrt{n}}$ is neglected and

$$y = \frac{1}{\sigma\sqrt{2\pi}}\ e^{-\frac{x^2}{2\sigma^2}}\left\{1 - \frac{\kappa}{2}\left(\frac{x}{\sigma} - \frac{x^3}{3\sigma^3}\right)\right\} \quad \ldots \ldots \quad (20)$$

when $\frac{1}{\sqrt{n}}$ is retained and $\frac{1}{n}$ neglected

These equations express the chances that when an n-fold experiment is made, as described above, the number of successes shall be x in excess of pn, where p is the chance of success in a single experiment.

The curve represented by $y = \dfrac{1}{c\sqrt{\pi}} e^{-\frac{x^2}{c^2}}$ is called the "normal curve of error." [*] Its shape is shown in the diagram at the end of the book.

An idea can be obtained of the importance of the term in $\dfrac{1}{\sqrt{n}}$ by taking $n = 1000$, $p = \frac{1}{10}$. Then $\sigma = \sqrt{90} = 9 \cdot 5$ and $\kappa = \cdot 084$ approx. The chance is sensibly affected when x is greater than σ.

When n is great the actual chance of one assigned number of successes is small, e.g. if $p = \frac{1}{2}$, $n = 1000$, the chance of exactly 500 (the most probable number of) successes is only $\frac{1}{40}$ approx. The measurement that we find useful, however, is not that of particular ordinates, but of the sum of the chances over a range of values, say from x_1 to x_2, where $x_2 - x_1$ is of the same order as $\sigma(= \sqrt{pqn})$.

By a well-known theorem † we can pass from summation of the ordinates to integration of an area, and the whole chance of a number of successes as great as $pn + x_1$ and not greater than $pn + x_2$ is $\int_{x_1}^{x_2} y\,dx$, where $y = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$ and terms involving $\dfrac{1}{\sqrt{n}}$ are neglected.

Writing z for $\dfrac{x}{\sigma}$, we have $\int_{x_1}^{x_2} y\,dx = \dfrac{1}{\sqrt{2\pi}} \int_{z_1}^{z_2} e^{-\frac{1}{2}z^2}\,dz$, and a table suitable for evaluating this function is given on p. 271.

In the following paragraph important constants connected with the function in question are obtained.

Area of curve $= \int_{-\infty}^{\infty} \dfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = $ limit of $(p + q)^n$ when n tends to infinity $= 1$.

$$\therefore \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2} dz = \sqrt{2\pi}; \quad \int_{-\infty}^{\infty} e^{-u^2}.\,du = \sqrt{\pi};$$

$$\int_{-\infty}^{\infty} e^{-au^2} du = \sqrt{\dfrac{\pi}{a}}; \quad \int_{-\infty}^{\infty} \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} dx = 1$$

---

[*] See Edgeworth, *Encyc. Brit.* Vol. XXII., article *Probability*, pp. 391 *seq.*
† Appendix, Note 4.

Write $m_s = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \cdot x^s dx$, for the sth moment about the average, which is the origin, the area being 1.

The curve is symmetrical about the ordinate through the origin, and $m_{2t+1} = 0$ for all values of t.[*]

$$m_2 = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-\frac{x^2}{2\sigma^2}} dx$$

$$= \left[-\frac{\sigma}{\sqrt{2\pi}} x e^{-\frac{x^2}{2\sigma^2}} \right]_{-\infty}^{\infty} + \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2}} dx$$

$$= 0 + \sigma^2 = \sigma^2 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (21)$$

as was already known from formula (13).

$$m_{2t} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x^{2t} e^{-\frac{x^2}{2\sigma^2}} dx$$

$$= \left[-\frac{\sigma}{\sqrt{2\pi}} x^{2t-1} e^{-\frac{x^2}{2\sigma^2}} \right]_{-\infty}^{\infty} + \frac{(2t-1)\sigma^2}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x^{2t-2} e^{-\frac{x^2}{2\sigma^2}} dx$$

$$= 0 + (2t-1)\sigma^2 m_{2t-2} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (22)$$

Hence $m_4 = 3\sigma^2 \cdot m_2 = 3\sigma^4 = 3m_2^2$, and

$$\kappa_2 = \beta_2 = \frac{m_4}{m_2^2} = 3, \quad i = \beta_2 - 3 = 0,$$

as may also be obtained from p. 264, when n is infinite.

$m_{2t} = (2t-1)(2t-3) \ldots 3 \cdot 1\sigma^{2t}$, by induction,

$$= \frac{(2t)!}{2^t \cdot t!} \sigma^{2t} \quad . \quad . \quad . \quad . \quad . \quad . \quad (23)$$

E.g. $m_6 = 15\sigma^6$, $m_8 = 105\sigma^8$.

η, the mean deviation (see p. 111), since the area is unity,

$$= \frac{2}{\sigma\sqrt{2\pi}} \int_0^{\infty} x e^{-\frac{x^2}{2\sigma^2}} dx = \left[-\frac{2\sigma}{\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \right]_0^{\infty} = \sigma \cdot \sqrt{\frac{2}{\pi}} . \quad . \quad (24)$$

and

$$\therefore \frac{\sigma}{\eta} = \sqrt{\frac{\pi}{2}}.$$

---

[*] For $m_{2t+1} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x^{2t+1} e^{-\frac{x^2}{2\sigma^2}} dx = \int_{-\infty}^{\infty} \phi(x) dx$, say,

$= \int_0^{\infty} \phi(x) dx + \int_{-\infty}^0 \phi(x) dx = \int_0^{\infty} \phi(x) dx - \int_0^{\infty} \phi(x') dx'$, where $x' = -x$, $= 0$.

The "probable error" (see p. 113) is obtained by finding from the table the value of z which makes $\int_0^z \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = \frac{1}{4}$. This value has been calculated as $x = z\sigma = \cdot 67449\sigma$.

A drawing of the curve is given at the end of the book. The points of inflection are obtained by equating $D_x^2 y$ to zero, where

$$ y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}. $$

Thus
$$ \log y + \text{const.} = -\frac{x^2}{2\sigma^2} $$

$$ \frac{1}{y} D_x y = -\frac{x}{\sigma^2} $$

$$ \frac{1}{y} \cdot D_{xy}^2 - \frac{1}{y^2}(D_x y)^2 = -\frac{1}{\sigma^2} $$

$\therefore$
$$ -\left(-\frac{x}{\sigma^2}\right)^2 = -\frac{1}{\sigma^2} \text{ at the points of inflection} $$

and
$$ x = \pm \sigma \quad \ldots \ldots \ldots (25) $$

The area of that part of the curve which stands on the base 0 to σ is, of course, the tabular value of

$$ \frac{1}{\sigma\sqrt{2\pi}} \int_0^\sigma e^{-\frac{1}{2}\frac{x^2}{\sigma^2}} dx = \frac{1}{\sqrt{2\pi}} \int_0^1 e^{-\frac{1}{2}z^2} dz = F(1) = \cdot 3413 ; $$

and by a similar use of the table we readily find the following approximate values :—

PROPORTION OF AREA OF CURVE STANDING ON CERTAIN BASES.

Base.	Area.	Base.	Area
0— ·2σ	·07926	— ·2σ to + ·2σ	·1585
0— ·6σ	·2257	+ ·2σ ,, + ·6σ	·1465
0— 1·0σ	·3413	+ ·6σ ,, + 1·0σ	·1156
0— 1·4σ	·4192	+ 1·0σ ,, + 1·4σ	·0779
0— 1·8σ	·4641	+ 1·4σ ,, + 1·8σ	·0449
0— 2·2σ	·4861	+ 1·8σ ,, + 2·2σ	·0220
0— 2·6σ	·4953	+ 2·2σ ,, + 2·6σ	·0092
0— 3·0σ	·49865	+ 2·6σ ,, + 3·0σ	·0033

NOTE.—The mean deviation and probable error are defined in Part I. pp. 111–3.

The mean deviation is the average without regard to sign of the differences between the measurements of the items which make the group and a central measurement (generally the arithmetic average).

The probable error is the distance which measured left and right from a central position includes exactly half the observations.

TABLE OF VALUES* OF $F(z) = \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_0^z e^{-\frac{1}{2}z^2}\,dz$

z	F(z)	z	F(z)	z	F(z)	z	F(z)	z	F(z)
·00	·0000	·50	·1915	1·00	·3413	1·50	·4332	2·00	·4772
·01	·0040	·51	·1950	1·01	·3438	1·51	·4345	2·02	·4783
·02	·0080	·52	·1985	1·02	·3461	1·52	·4357	2·04	·4793
·03	·0120	·53	·2019	1·03	·3485	1·53	·4370	2·06	·4803
·04	·0160	·54	·2054	1·04	·3508	1·54	·4382	2·08	·4812
·05	·0199	·55	·2088	1·05	·3531	1·55	·4394	2·10	·4821
·06	·0239	·56	·2123	1·06	·3554	1·56	·4406	2·12	·4830
·07	·0279	·57	·2157	1·07	·3577	1·57	·4418	2·14	·4838
·08	·0319	·58	·2190	1·08	·3599	1·58	·4429	2·16	·4846
·09	·0359	·59	·2224	1·09	·3621	1·59	·4441	2·18	·4854
·10	·0398	·60	·2257	1·10	·3643	1·60	·4452	2·20	·4861
·11	·0438	·61	·2291	1·11	·3665	1·61	·4463	2·22	·4868
·12	·0478	·62	·2324	1·12	·3686	1·62	·4474	2·24	·4875
·13	·0517	·63	·2357	1·13	·3708	1·63	·4484	2·26	·4881
·14	·0557	·64	·2389	1·14	·3729	1·64	·4495	2·28	·4887
·15	·0596	·65	·2422	1·15	·3749	1·65	·4505	2·30	·4893
·16	·0636	·66	·2454	1·16	·3770	1·66	·4515	2·32	·4898
·17	·0675	·67	·2486	1·17	·3790	1·67	·4525	2·34	·4904
·18	·0714	·68	·2517	1·18	·3810	1·68	·4535	2·36	·4909
·19	·0753	·69	·2549	1·19	·3830	1·69	·4545	2·38	·4913
·20	·0793	·70	·2580	1·20	·3849	1·70	·4554	2·40	·4918
·21	·0832	·71	·2611	1·21	·3869	1·71	·4564	2·42	·4922
·22	·0871	·72	·2642	1·22	·3888	1·72	·4573	2·44	·4927
·23	·0910	·73	·2673	1·23	·3907	1·73	·4582	2·46	·4931
·24	·0948	·74	·2703	1·24	·3925	1·74	·4591	2·48	·4934
·25	·0987	·75	·2734	1·25	·3944	1·75	·4599	2·50	·4938
·26	·1026	·76	·2764	1·26	·3962	1·76	·4608	2·52	·4941
·27	·1064	·77	·2794	1·27	·3980	1·77	·4616	2·54	·4945
·28	·1103	·78	·2823	1·28	·3997	1·78	·4625	2·56	·4948
·29	·1141	·79	·2852	1·29	·4015	1·79	·4633	2·58	·4951
·30	·1179	·80	·2881	1·30	·4032	1·80	·4641	2·60	·4953
·31	·1217	·81	·2910	1·31	·4049	1·81	·4649	2·62	·4956
·32	·1255	·82	·2939	1·32	·4066	1·82	·4656	2·64	·4959
·33	·1293	·83	·2967	1·33	·4082	1·83	·4664	2·66	·4961
·34	·1331	·84	·2995	1·34	·4099	1·84	·4671	2·68	·4963
·35	·1368	·85	·3023	1·35	·4115	1·85	·4678	2·70	·4965
·36	·1406	·86	·3051	1·36	·4131	1·86	·4686	2·72	·4967
·37	·1443	·87	·3078	1·37	·4147	1·87	·4693	2·74	·4969
·38	·1480	·88	·3106	1·38	·4162	1·88	·4699	2·76	·4971
·39	·1517	·89	·3133	1·39	·4177	1·89	·4706	2·78	·4973
·40	·1554	·90	·3159	1·40	·4192	1·90	·4713	2·80	·4974
·41	·1591	·91	·3186	1·41	·4207	1·91	·4719	2·82	·4976
·42	·1628	·92	·3212	1·42	·4222	1·92	·4726	2·84	·4977
·43	·1664	·93	·3238	1·43	·4236	1·93	·4732	2·86	·4979
·44	·1700	·94	·3264	1·44	·4251	1·94	·4738	2·88	·4980
·45	·1736	·95	·3289	1·45	·4265	1·95	·4744	2·90	·4981
·46	·1772	·96	·3315	1·46	·4279	1·96	·4750	2·92	·4982
·47	·1808	·97	·3340	1·47	·4292	1·97	·4756	2·94	·4984
·48	·1844	·98	·3365	1·48	·4306	1·98	·4761	2·96	·4985
·49	·1879	·99	·3389	1·49	·4319	1·99	·4767	2·98	·4986

z	F(z)	z	F(z)	z	F(z)
3·00	·49865	3·60	·499841	4·50	·499997
3·20	·49931	3·80	·499928		
3·40	·49966	4·00	·499968		

* Based on Dr. Sheppard's 7 figure Tables, *Biometrika*, Vol. II, Part II.

It has been calculated that $F(z) = \frac{1}{4}$ when $z = \cdot 67449$ approx. The quartiles of the curve are therefore at

$$\pm \cdot 67449\sigma \quad . \quad . \quad . \quad . \quad . \quad (26)$$

and it is just as likely as not that a single observation will be within this range as without it. $\cdot 67449\sigma$ is therefore the " probable error " and is frequently used in preference to σ to measure precision.

Algebraic Chance and Experience.

The analysis so far has been purely abstract, the illustrations from cards and dice only having been taken to visualise the phrase " equally likely." We must now consider what evidence there is that successes do occur in proportion to their algebraic probability. Though we should certainly be surprised if, in simple cases, successes were in a different proportion—if, for example, we found that 90 out of 100 coins tossed fell head uppermost, or 50 repeated draws of one card from a complete pack (shuffled after replacing each card drawn) were all hearts—yet this feeling hardly gives more than a presumption that in the universe there is some method in apparently chance events. We must appeal to experience and experiment. In a general way, it is the experience of players of games of chance that events do happen at any rate roughly in proportion to their algebraic probabilities ; canons of correct play in whist were based on this, and the odds were given in accordance with calculated probability. Insurance, both accident and life, is based on the belief that events in the bulk are predictable, though individual occurrences appear to be fortuitous, and this belief has been continually justified. A great number of experiments have been carried out directly for the purpose of comparing the frequency of the occurrence of events with their *a priori* chances, with very marked success. We can never, however, obtain a certainty that the preliminary condition of equal probability is satisfied completely, nor can we expect to obtain more than an approximate verification.

Rough experiments can easily be made by quite simple means.

Thus from numerous packs of cards, from which the picture cards had been removed, 4 cards were drawn and the total of

the pips on them counted and then the cards replaced., This was done 90 times.

The chance of getting a total of r pips, if the number of packs was so large that the draws of the separate cards in the quartets could be taken as independent, is the coefficient of x^r in $\frac{1}{10^4}(x + x^2 + \ldots + x^{10})^4$, *i.e.* in $\frac{x^4}{10^4} \cdot \left(\frac{1 - x^{10}}{1 - x}\right)^4$, and may be tabulated with the results of the experiment as follows :—

	Aggregate chance.	X 90= " Expectation."	Experimental result.
$r =$ ·4 to 9	·0126	1·134	0
10 ,, 14	·0871	7·839	7
15 ,, 19	·2375	21·375	27
20 ,, 24	·3256	29·304	25
25 ,, 29	·2375	21·375	22
30 ,, 34	·0871	7·839	9
35 ,, 40	·0126	1·134	0
	1·0000	90·000	90

The total of all the pips in the 90 quartets was 1956, and the average per card 5·43. The average on all the cards in the packs was 5·5.

It is evidence that the experiment corresponds with the expectation, approximately at any rate.

Bernoulli's Laws.

We must next inquire what correspondence between theoretical and expected frequency the theory itself leads us to expect. The Law of Error supplies a test.

Consider the group $r = 15$ to 19 in the above experiment. The chance of finding a number in this range is ·2375 $= p$. In 90 experiments the chance of finding a number in this range t times is the $\overline{t + 1}^{th}$ term of $(q + p)^{90}$. The most likely number of successes is 21 or 22 and the standard deviation of the possible number of successes is $\sqrt{pqn}$ where $n = 90$, *i.e.*, about 4. In such a multiple experiment many times repeated, the chance of getting anything from 17 to 26 successes in the group is found from the Table to be about $\frac{2}{3}$; that we should obtain so great a number as 27 (as in the experiment tabulated) the chance is about $\frac{1}{6}$. It is very unlikely that we should have a divergence from 21 by as much as 3 times the standard deviation ; that is, more than 33 or less than 9 occurrences are very improbable.

T* ,

This process, stated more generally, leads to Bernoulli's Laws, which may be paraphrased as follows. If an experiment, in which the chance of success is p, is performed n times, and $p'n$ is written for the number of successes, then as n is increased p' tends to approach p. The chance of the occurrence of a deviation greater than $p \sim p'$, is $2 \int_z^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$,

where $$z = \frac{p \sim p'}{\sqrt{\dfrac{p(1-p)}{n}}},$$

and hence as $\sqrt{n}$ increases the chance of any assigned deviation diminishes. By increasing n sufficiently the chance can be made as small as we please.*

Now it is the result of general experience and many experiments that Bernoulli's Laws can be realised in fact.

If, then, we can obtain the condition of *a priori* equally likely occurrences, we may calculate the chances of various events by the methods of mathematical probability, and expect that our calculations will be realised in fact within a margin determinable by the law of error.

On the following pages the results of various experiments are shown. The first three compare the distribution found with that given by the law of error, and the remainder show the working method of determining the size of a class in a large group by the method of sampling.

Examples.

1. If a digit is taken at random the chance that it will be less than 5 (0, 1, 2, 3 or 4) is $\frac{1}{2}$. The digits in the 7th decimal place of a book of logarithms were taken 50 at a time and the number (r) of digits less than 5 was noted. The chance of finding r such digits is the $\overline{r+1}^{th}$ term in the expansion of $(\frac{1}{2} + \frac{1}{2})^{50}$. $n = 50$, $p = q = \frac{1}{2}$, $\sqrt{pqn} = 3 \cdot 535 = \sigma$.

pn, the most probable number, is 25. The chance of not exceeding $25 + x$ is $F(z)$ in the table, p. 271, where $z = \dfrac{x}{\sigma} = \dfrac{x}{3 \cdot 535}$, if we

---

* Notice that $p \sim p'$ is the deviation of the proportions. The resulting actual deviation is $pn \sim p'n$, and z should then be written

$$\frac{pn \sim p'n}{\sqrt{p(1-p)n}},$$

and the chance *increases* as $\sqrt{n}$ increases.

assume that $n = 50$ is large enough in a symmetrical curve to allow the use of the normal curve instead of the binomial series.

The 50-fold experiment was performed 300 times.

r	z	F(z)	Differences ×300=		Expected† number	Actual of occurrences.		
13.5	− 3·2522	·4994						
14.5	− 2·9694	·4986	·0008	·2	0	1	at	14
15.5	− 2·6866	·4966	·0020	·6	0 or 1	0	„	15
16.5	− 2·4038	·4919	·0047	1·4	1 or 2	3	„	16
17.5	− 2·1210	·4831	·0088	2·6	2 or 3	2	„	17
18.5	− 1·8382	·4670	·0161	4·8	5	3	„	18
19.5	− 1·5554	·4400	·0270	8·1	8	7	„	19
20.5	− 1·2726	·3984	·0416	12·5	12 or 13	9	„	20
21.5	− ·9898	·3389	·0595	17·85	18	18	„	21
22.5	− ·7070	·2602	·0787	23·6	24	26	„	22
23.5	− ·4242	·1643	·0950	28·8	29	21	„	23
24.5	− ·1414	·0561 ⎫	·1082	32·5	32 or 33	32	„	24
25.5	+ ·1414	·0561 ⎭	·1122	33·7	34	42	„	25
26.5	+ ·4242	·1643	·1082	32·5	32 or 33	36	„	26
27.5	+ ·7070	·2602	·0959	28·8	29	30	„	27
28.5	+ ·9898	·3389	·0787	23·6	24	28	„	28
29.5	1·2726	·3984	·0595	17·85	18	15	„	29
30.5	1·5554	·4400	·0416	12·5	12 or 13	16	„	30
31.5	1·8382	·4670	·0270	8·1	8	5	„	31
32.5	2·1210	·4831	·0161	4·8	5	2	„	32
33.5	2·4038	·4919	·0088	2·6	2 or 3	2	„	33
34.5	2·6866	·4966	·0047	1·4	1 or 2	1	„	34
35.5	2·9694	·4986	·0020	·6	0 or 1	1	„	35
36.5	3·2522	·4994	·0008	·2	0	0	„	36

299·6

The agreement is as close as the theory leads us to expect (see Chapter X). The standard deviation *a priori* is $\sqrt{pqn} = 3·535$. We can also find the standard deviation of the observations *a posteriori* by taking the square root of the second moment as on p. 253. The average is 25·043. The second moment of the observations about an origin at 25 is $(1 \times 11^2 + 0 \times 10^2 + 3 \times 9^2 + \ldots + 1 \times 10^2) \div 300 = 11·30$, and about the average is $11·300 − ·043^2 = 11·298$. The square root is 3·361, which differs from the *a priori* value by ·174, which is a not improbable deviation (see formula (120) below).

2. Instead of finding the expectation at each value, we can test the distribution by the method illustrated in the following example.

In a book, in which a page contained 37 lines, it was counted on each of 100 pages in how many cases the first (complete)

* Thus when $r = 13·5$ and $14·5$, $F(z) = ·4994$ and $·4986$. The difference ·0008, × 300, is the expected number at $r = 14$.

† Nearest whole numbers from the previous column.

T* 2.

word in a line contained 1, 2, or 3 letters. In 3700 lines such first words occurred 1317 times. The chance, then, that a first word contained 3 letters or less was。

$$p = \tfrac{1317}{3700}, \; q = 1 - p = \tfrac{2383}{3700}.$$

The chance of finding r such first words in a page was approximately the $\overline{r + 1}^{\text{th}}$ term in $(q + p)^{37}$.

The *a priori* standard deviation is $\sqrt{pqn} = 2 \cdot 913 = \sigma$.

The occurrences were as follows.

Number of first words of 3 letters or less.	Number of pages on which these occurred.
7	1
8	2
9	9
10	6
11	8
12	17
13	15
14	12
15	13
16	5
17	4
18	2
19	3
20	2
21	0
22	1

Average $13 \cdot 17 = \bar{x}$; standard deviation calculated from the observations $2 \cdot 90$. Now calculate the number of cases to be expected in grades each of σ measured from the average.

		Difference × 100.	Occurrences.	
$\bar{x} - 3\sigma = 4 \cdot 47$	$F(-3) = \cdot 499$		Under $7\frac{1}{2}$	1
$\bar{x} - 2\sigma = 7 \cdot 37$	$F(-2) = \cdot 477$	$2 \cdot 3$	$7\frac{1}{2}$ to $10\frac{1}{2}$	17
$\bar{x} - \sigma = 10 \cdot 27$	$F(-1) = \cdot 341$	$13 \cdot 6$	$10\frac{1}{2}$,, $13\frac{1}{2}$	40
$\bar{x} = 13 \cdot 17$	$F(0) = \cdot 0$	$34 \cdot 1$	$13\frac{1}{2}$,, $16\frac{1}{2}$	30
$\bar{x} + \sigma = 16 \cdot 07$	$F(1) = \cdot 341$	$34 \cdot 1$	$16\frac{1}{2}$,, $19\frac{1}{2}$	9
$\bar{x} + 2\sigma = 18 \cdot 97$	$F(2) = \cdot 477$	$13 \cdot 6$	$19\frac{1}{2}$,, $22\frac{1}{2}$	3
$\bar{x} + 3\sigma = 21 \cdot 87$	$F(3) = \cdot 499$	$2 \cdot 3$		
		$100 \cdot 0$		100

In observations where the measurements are necessarily integral, it is not easy to adjust the grades to multiples of σ. But where the observational grades are narrow, or the measurements continuous, this method (proceeding by equal submultiples of σ) is rapid, and since the grading can be decided before the test is applied, affords a good and simple test.

3. A similar experiment was made with a list of firms, in which there were 74 pages containing about 40 names each. Each

firm had been marked for administrative purposes if it employed a certain number of women. One-fifth of all the firms were so marked. On any page the chance of finding r firms was therefore the $r + 1^{\text{th}}$ term in $(q + p)^{40}$ where

$$p = \tfrac{1}{5}, \ \sigma = \sqrt{(\tfrac{1}{5} \cdot \tfrac{4}{5} \cdot 40)} = 2 \cdot 53, \ pn = 8.$$

								Expected.	Actual.
Between	$pn+2\sigma$	and	$pn+3\sigma$	.	.	.	.	$1\cdot7$	2 or 3
,,	$+\tfrac{2}{3}\sigma$	,,	$+2\sigma$	.	.	.	.	$3\cdot3$	5, 6 or 7
,,	$+\sigma$	,,	$+\tfrac{2}{3}\sigma$	.	.	.	.	$6\cdot8$	4 or 5
,,	$+\tfrac{1}{2}\sigma$	,,	$+\sigma$	.	.	.	.	$11\cdot0$	9, 10, 11
,,	0	,,	$+\tfrac{1}{2}\sigma$	.	.	.	.	$14\cdot2$	13 or 14
,,	$-\tfrac{1}{2}\sigma$	,,	$+0$	.	.	.	.	$14\cdot2$	15 or 16
,,	$-\sigma$	,,	$-\tfrac{1}{2}\sigma$	.	.	.	.	$11\cdot0$	8
,,	$-\tfrac{2}{3}\sigma$	,,	$-\sigma$	.	.	.	.	$6\cdot8$	7
,,	-2σ	,,	$-\tfrac{2}{3}\sigma$	.	.	.	.	$3\cdot3$	2
,,	-3σ	,,	-2σ	.	.	.	.	$1\cdot7$	5

The alternatives in the final column are due to the difficulty of adjusting the entries to the predetermined grades.

In this case the preliminary condition of independence is not completely fulfilled ; the chance of finding a marked name should not be affected by the presence or absence of marked names on the same page ; but in fact in some cases the name of a firm was repeated for each of its branches, and all the branches did or all did not employ women.

Application to Sampling.

One of the principal uses of the theorem relating to the number of successes to be expected in a given number of trials is in the examination of a large group by means of samples. In its simplest form the method is as follows.

In a " universe " containing N things or persons, pN possess a defined attribute, where N is known but p is not known.

n things are selected at random from the universe, and of them $p'n$ are found to possess the attribute.

If $\dfrac{n}{N}$ is small,* and if in the process of selection everything in the universe has an equal chance of being chosen, and if the choice of one thing does not influence the choice of any other, then the chance of finding $(p + x)n$ things is given by

* The necessary correction, when $\dfrac{n}{N}$ is not negligible, is given below, pp. 282–4.

$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\frac{x^2}{\sigma^2}}$, where $\sigma = \sqrt{\dfrac{pq}{n}}$, and the table on p. 271 can be applied. The precision, measured by $\dfrac{1}{\sigma}$, increases with $\sqrt{n}$.

It is shown below (p. 417) that in evaluating σ, the value p', observed in the sample, can be substituted for the unknown true value p.

The result may be stated thus: the value of p in the universe is $p' \pm \sqrt{\dfrac{p'(1 - p')}{n}}$, the expression meaning that p' is the most probable value from the data, and that the chances of variations from p' are given by the Table, p. 271, where the standard deviation (the unit in the Table) is $\sqrt{\dfrac{p'(1 - p')}{n}}$.

It is clear that this value can only be applied to the defined universe, the members of which have the chance of being enumerated. The importance of this and other conditions can be best illustrated by an example.

In Reading 609 working-class houses were visited, and in 154 of them it was found that there were more than 1 and less than 2 inhabitants per room. $n = 609$, $p'n = 154$, $p' = \cdot253$, $\sqrt{p'q'/n} = \cdot0176$. The proportion of houses thus occupied is $\cdot253 \pm \cdot0176$.

The " universe " here is the group of houses (about 12,000) from which the 609 were selected. This group was determined from a local directory, from which middle-class and large houses were eliminated by the help of a list of " principal residents " and by local knowledge, and non-residential houses were omitted. The accuracy of the measurement for working-class Reading depends on the completeness and accuracy of the directory and on the appositeness of the method of elimination. If a rookery of slum dwellings had been omitted, by so much the universe would have been curtailed ; or if a street of middle-class houses had been included the universe would have been extended, unless in the process of investigation the error had been found.

In this case the selection was made by marking one house in 20 throughout the amended directory. It is shown on p. 332 that this gives a more precise result than if a purely random method had been followed. A general method of

securing randomness is to give numbers from 1 to N to the things in the universe, and by the use of tables of figures or otherwise select n numbers.* Great care must be taken to ensure pure randomness or some method which gives a more precise result than pure randomness. It was found, for example, that in the latitude experiment (p. 281) randomness was not obtained by selecting pages and dropping a pencil on names; the entries in a page were not independent of each other. Any divergence from the rule that every item must have the same chance of inclusion may affect the result disastrously.

Of course inaccuracy of information (*e.g.,* as to the number of persons resident in a house) is to be avoided; but if the errors due to this source are equally likely to be in excess or defect, the result is not much affected.

It should be noticed that the accuracy of the result depends on n the number in the sample, and not on N the number in the universe. The size of the universe only affects the problem in that, when the N things are numerous and scattered, it is difficult to get an accurate enumeration and secure that each has an equal chance of being chosen, and it becomes possible that parts are omitted from ignorance of their existence, which differ essentially from the major parts included. Further when p is small, pN may be moderately large, while pn is relatively small. Now if pn is small, the approximation to the curve of error (p. 265) tends to break down, and the term involving $\kappa \left(\dfrac{1 - 2p}{\sqrt{p\,(1 - p)n}} \right)$ is not negligible; so that the terms of the binomial $(q + p)^n$ should be used instead of the integral table. A little examination of numerical cases will show that for certain small values of p it is quite possible that no thing having the attribute will be found; thus, if 30 houses in a town containing 10,000 houses are overcrowded, and 800 houses are examined, the chance of finding no overcrowded house is $q^n p^0$, where $p = \cdot003$, $q = \cdot997$, $n = 800$, that is $\cdot09$; so that a report based on the sample might not contain reference to overcrowding, unless to say that there was no evidence of it.

---

* For example, if N = 10,000 and n = 500, we might take the last four digits of pages in 7 figure tables till we had 500 numbers all between 0 and 10,001, and investigate the things to which these numbers were affixed. This method was used in the experiment on the number of persons in a parish. See next page.

But if $p = \cdot 03$, $q^n p^0$ is only about $\dfrac{3}{10^{11}}$, and some instances would certainly be found. As to the chances of occurrence of small numbers, see p. 284 below.

Finally it should be emphasised that when the things that should be included are determined by marking in a list ōr otherwise, no difficulties of measurement should be allowed to stand in the way of their inclusion. If a householder refuses information, or part of a consignment of goods is out of the way, there is a presumption that the characteristics of the house or the goods are not normal, and unless the difficulty is overcome, some part of the universe is not represented.

Examples of Sampling.

1. The 12,830 civil parishes enumerated in the Census of England and Wales, 1911, were numbered, and 250 selected by numbers taken from logarithmic tables. The following table compares the distribution of the parishes according to their populations in the sample and in the whole group (which is set out in the Census Volume, Cd. 6258, p. 428).

NUMBER OF PERSONS IN PARISH.

	Under 100.	100 to 200.	200 to 300.	300 to 400.	400 to 500.	500 to 1000.	1000 or more.
Number of parishes in sample of 250.	35	52	42	27	20	41	33
1000 p'	140	208	168	108	80	164	132
1000 $\sqrt{\dfrac{p'q'}{250}}$	22	26	24	20	17	23	21
Actual per 1000	152	192	147	108	80	173	146

Here (to take the first column as an example) 35 were found in the sample of 250 with population less than 100.

$$p' = \frac{35}{250} = \cdot 14.$$

The forecast per 1000 parishes is therefore $\cdot 14$ of $1000 = 140$. The standard deviation of p' is $\sqrt{\dfrac{p'(1 - p')}{250}} = \cdot 022$, p. 278, and therefore the standard deviation of $1000p'$, i.e. of the forecast 140, is 22. Actually in England and Wales there were 152 per 1000 parishes with less than 100 people. The forecast differs from the fact by about half the standard deviation. (*Statistical Journal*, 1912–13, p. 182.)

2. From a list of the rates of dividends of 3878 companies 400 were selected and tabulated.

RATE OF DIVIDEND PER CENT.

	Below £3	£3	£4	£5	£6	£8
Number of companies in sample	34	108	117	60	48	33
1000 p'	85	270	292½	150	120	82½
1000 $\sqrt{\dfrac{p'q'}{400}}$	14	22	23	18	16	14
In full list per 1000 . . .	75	272	311	177	108	57

(*Statistical Journal*, 1906, p. 552.)

3. From a geographical index containing 31,210 names 500 places were selected and their latitudes tabulated. To secure randomness the columns of names were numbered and selection made from numbers in mathematical tables ; a foot-rule was placed over the column, and the entry against the number of inches on the rule determined by the first digit of the longitude of the first place in the column was selected. This elaborate method was found necessary to secure independence.

LATITUDE, NORTH OR SOUTH.

	0° to 10°	10° to 20°	20° to 30°	30° to 40°	40° to 50°	50° to 60°	60° to 70°	70° to 80°	80° to 90°
Number of places in sample	22	56	104	103	93	112	9	1	0
1000 p' . . .	44	112	208	206	186	224	18	2	0
1000 $\sqrt{\dfrac{p'q'}{500}}$ . . .	9	14	18	18	17	19	6	?	?
In full list per 1000 .	51	111	201	200	200	215	18	3·4	0·9

Notice that the places north of 80° N. and south of 80° S. were missed in the accident of the selection. In another selection where $n = 2000$, 1 per 1000 were found in these latitudes.

4. Out of the householders' schedules of the 1911 Census, 1 in 50 in order throughout the files were selected in Shoreditch, and the *personnel* of the households classified.

	OCCUPIED PERSONS.				UNOCCUPIED.		
	Males.		Females.		—		Total.
	Over 20 years.	Under 20 years.	Over 18 years.	Under 18 years.	Over 14 years.	Under 14 years.	
Number of persons in sample . . .	538	112	310	74	386	718	2138
1000 p' . . .	251	52	145	35	181	336	1000
1000 $\sqrt{\dfrac{p'q'}{2138}}$ . .	9	5	8	4	8	10	—
Distribution per 1000 from Census tables	258	55	144	33	185	325	1000

Case when the Universe is not practically Unlimited or the Selections are not Independent.

In the statement of the experiment which leads to the normal curve of error (pp. 263 *seq.*) it was assumed that the chance of success for each throw or draw was always the same (p. 261), and that each trial was uninfluenced by what had already happened. In practice this condition is seldom completely satisfied, but we can prove in a similar manner that the normal law of error is obtained under a wider hypothesis.*

Let a universe contain N objects, of which pN possess a certain quality or attribute, and qN do not ($p + q = 1$). Let a selection of n be made in such a way that every object in the universe has the same chance of being chosen. Write P_x for the probability that $pn + x$ of the selected objects shall possess the quality in question. *E.g.*, if the "universe" is a box containing 1000 balls of which 100 are white and the rest coloured, and if the contents are thoroughly mixed and 50 selected, then N = 1000, $p = \frac{1}{10}$ (where *white* is the attribute), $n = 50$, $pn = 5$, and P_x is the probability that $5 + x$ white balls are present in the selection.

The whole number of different possible selections is $_NC_n$.

The number of selections in which $pn + x$ are white and the remainder $(qn - x)$ are coloured is $C_{pn+x} \times {}_{qN}C_{qn-x}$.

Hence $P_x = \dfrac{_{pN}C_{pn+x} \times {}_{qN}C_{qn-x}}{_NC_n}$ *

$$= \frac{(pN)!\,(qN)!\,n!\,M!}{(pn+x)!\,(pM-x)!\,(qn-x)!\,(qM+x)!\,N!},$$

where $M = N - n$.

Apply Stirling's theorem to the factorials, neglecting $\dfrac{1}{pn}, \dfrac{1}{n}$ and smaller quantities. (App. formula (134).)

$$P_0 = (pN)^{pN}(qN)^{qN}(n)^n M^M (pn)^{-pn}(pM)^{-pM}(qn)^{-qn}(qM)^{-qM} N^{-N}$$
$$\cdot (2\pi)^{\frac{1}{2}-\frac{1}{2}} \cdot e^0 \cdot \left(\frac{pNqNnM}{pnpMqnqMN}\right)^{\frac{1}{2}},$$

the index of e being

$$pn + pM + qn + qM + N - pN - qN - n - M$$
$$= 0, \text{ since } p + q = 1.$$

---

* *E.g.* the chance of obtaining 3 aces in a hand of 13 dealt from a pack of 52 is $P_2 = {}_4C_3 \times {}_{48}C_{10} \div {}_{52}C_{13}$; here N = 52, $n = 13$, $p = \frac{1}{13}$, $x = 2$. $P_2 = \cdot 041$ approx.

When the indices are collected it is found that

$$P_0 = \left(\frac{N}{2\pi pqn M}\right)^{\frac{1}{2}} \quad \ldots \ldots \ldots \ldots \ldots \ldots \ldots \quad (27)$$

$$\frac{P_x}{P_0} = \frac{(pn)!\,(pM)!\,(qn)!\,(qM)!}{(pn+x)!\,(pM-x)!\,(qn-x)!\,(qM+x)!}$$

$$= \frac{(pn)^{pn}(pM)^{pM}(qn)^{qn}(qM)^{qM}\cdot(2\pi)^0\cdot e^0\cdot(pn\cdot pM\cdot qn\cdot qM)^{\frac{1}{2}}}{(pn+x)^{pn+x+\frac{1}{2}}(pM-x)^{pM-x+\frac{1}{2}}(qn-x)^{qn-x+\frac{1}{2}}(qM+x)^{qM+x+\frac{1}{2}}}$$

$$\therefore \frac{P_0}{P_x} = \left(1+\frac{x}{pn}\right)^{pn+x+\frac{1}{2}}\cdot\left(1-\frac{x}{pM}\right)^{pM-x+\frac{1}{2}}\cdot\left(1-\frac{x}{qn}\right)^{qn-x+\frac{1}{2}}\cdot\left(1+\frac{x}{qM}\right)^{qM+x+\frac{1}{2}}$$

$$\log P_x/P_0 = -\left(pn+x+\tfrac{1}{2}\right)\log\left(1+\frac{x}{pn}\right) - \left(qn-x+\tfrac{1}{2}\right)\log\left(1-\frac{x}{qn}\right)$$

$$\qquad -\left(pM-x+\tfrac{1}{2}\right)\log\left(1-\frac{x}{pM}\right) - \left(qM+x+\tfrac{1}{2}\right)\log\left(1+\frac{x}{qM}\right)$$

$$= -\left(pn+x+\tfrac{1}{2}\right)\left(\frac{x}{pn}-\frac{x^2}{2p^2n^2}+\ldots\right)$$

$$\qquad +\left(qn-x+\tfrac{1}{2}\right)\left(\frac{x}{qn}+\frac{x^2}{2q^2n^2}+\ldots\right)$$

$$\qquad +\left(pM-x+\tfrac{1}{2}\right)\left(\frac{x}{pM}+\frac{x^2}{2p^2M^2}+\ldots\right)$$

$$\qquad -\left(qM+x+\tfrac{1}{2}\right)\left(\frac{x}{qM}-\frac{x^2}{2q^2M^2}+\ldots\right)$$

$$= \frac{x}{2}\left(-\frac{1}{pn}+\frac{1}{qn}+\frac{1}{pM}-\frac{1}{qM}\right) - \frac{x^2}{2}\left(\frac{1}{pn}+\frac{1}{qn}+\frac{1}{pM}+\frac{1}{qM}\right)$$

$$\qquad +\frac{x^2}{4}\left(\frac{1}{p^2n^2}+\frac{1}{q^2n^2}+\frac{1}{p^2M^2}+\frac{1}{q^2M^2}\right)+\ldots\quad .$$

n is of course less than N, and we may take it (without loss of generality) as less than $\frac{1}{2}$N and therefore less than M.

Let pn and qn be at least moderately large, so that we proceed in ascending powers of $\dfrac{1}{\sqrt{pn}}$.

A solution is then obtained if we take $\dfrac{x^-}{n}$ as a quantity comparable with unity, and therefore $\dfrac{x}{n}$ as of order $\dfrac{1}{\sqrt{n}}$ and $\dfrac{x^2}{n^2}$ as of order $\dfrac{1}{n}$, as on p. 266 above.

Then neglecting terms of orders $\dfrac{1}{\sqrt{n}}$ and higher, we have

$$\log P_x/P_0 = -\frac{x^2}{2}\Big(\frac{p+q}{pqn}+\frac{p+q}{pqM}\Big) = -\frac{x^2\,(x+M)}{2pqnM} = -\frac{x^2N}{2pqnM}\cdot$$

Write σ^2 for $\dfrac{pqnM}{N}$.

$$P_0 = \frac{1}{\sigma\sqrt{2\pi}}$$

and
$$y = P_x = \frac{1}{\sigma\sqrt{2\pi}}\cdot e^{-\frac{x^2}{2\sigma^2}}$$

This is the normal curve of error, and σ (as above shown, formula (21)) is its standard deviation.

$$\sigma^2 = pqn\cdot\frac{N-n}{N} = pqn\Big(1-\frac{n}{N}\Big) \quad \cdot\ \cdot\ \cdot \quad (28)$$

and is smaller than its value (pqn) under the conditions of pp. 261–7, but tends to reach it (as it should) when N becomes indefinitely great.

Law of Small Numbers.

In the deduction of the normal curve from the terms of $(p+q)^n$ it was assumed that not only n, but also pqn, was large. An interesting case arises when p is so small that pn is no longer large, q being in that case nearly equal to 1.

Let $u = pn$, and be a small finite number.

$$p = \frac{u}{n}, \quad q = 1 - \frac{u}{n}.$$

The chance of r successes in n independent experiments is

$$P_r = \frac{n!}{(n-r)!\,r!}p^r q^{n-r}$$

$$= \Big(1-\frac{1}{n}\Big)\Big(1-\frac{2}{n}\Big)\ldots\Big(1-\frac{r-1}{n}\Big)\cdot\frac{u^r}{r!}\cdot\Big(1-\frac{u}{n}\Big)^n q^{-r}$$

Neglect $\dfrac{1}{n}$; then the product of the $\overline{r-1}$ factors in brackets, which is between 1 and $1 - \dfrac{r(r-1)}{2n}$, may be taken as 1.

Also $\left(1 - \dfrac{u}{n}\right)^n$ tends to e^{-u}, and

$$q^{-r} = \left\{\left(1 - \dfrac{u}{n}\right)^{-n}\right\}^{\frac{r}{n}} \text{ tends to } \left(e^u\right)^{\frac{r}{n}}$$

and to 1, as $\dfrac{r}{n}$ tends to o.

In all $\qquad\qquad\quad P_r = e^{-u} \cdot \dfrac{u^r}{r\,!} \qquad$ . . . . . . (29)

when $\dfrac{1}{n}, \dfrac{r}{n}$ and $\dfrac{r^2}{n}$ are neglected.

$$\sigma^2 = pqn = u\left(1 - \dfrac{u}{n}\right) \qquad \therefore \sigma = \sqrt{u}, \text{ approx.} \quad \text{. . . .} \quad (30)$$

$$\kappa = \dfrac{q-p}{\sigma} = \left(1 - \dfrac{2u}{n}\right) \Big/ \sqrt{\left\{u\left(1 - \dfrac{u}{n}\right)\right\}} = \dfrac{1}{\sqrt{u}}, \text{ approx.} \text{ . } (31)$$

The whole curve is then determined by u, without separate reference to p and n, since its average is u, its standard deviation $\sqrt{u}$, and its "κ" $\dfrac{1}{\sqrt{u}}$. It follows that the values of p and n are not easily determined separately from observations.

The greatest term of the binomial expansion is

$$P_{pn} = \dfrac{e^{-u}u^u}{u\,!}, *$$

when u is integral, and then

$$P_r = P_{pn} \cdot \dfrac{u^{r-u} \cdot u\,!}{r\,!} = \dfrac{P_{pn}}{\dfrac{r}{u}\left(\dfrac{r}{u} - \dfrac{1}{u}\right) \cdots \left(\dfrac{r}{u} - \dfrac{r-u-1}{u}\right)},$$

and this rapidly becomes small as $\dfrac{r}{u}$ passes through integral values. $E.g.$ if $u = 6$, and $r = 3u$, $P_{3u} = \cdot00004$.

Consequently the observed values never differ greatly from their average. Attention has been directed to the agreement between the fluctuations of small numbers and the law of distribution thus described, and examples have been given by Bortkiewicz (*Das Gesetz der kleinen Zahlen*, 1898) and Mortara (*Annali di Statistica*, Serie V, vol. 4, 1912). It is

---

* If u is as great as 10, this differs from $\dfrac{1}{\sqrt{2\pi u}}$ by less than 1 per cent.

also interesting to notice that the theory leads to what may be called the *permanence of small numbers.* If among a great number of things there are a few which present some particular feature, it is a matter of common experience that this small number is seldom much exceeded and seldom entirely vanishes ; this experience applies to accidents, fires, the traditional " Derby dog," and to the rare events and coincidences with which some newspapers fill their columns. Specialists in all professions, from the doctor who treats only one obscure disease of the ear to the dealer in curiosities, make their livelihood dependent on this permanence of small numbers.

To take an example : Out of some 530,000 deaths annually from all causes the following are the numbers from splenic fever in the years 1875 to 1894 :—

5, 4, 10, 14, 12, 18, 9, 15, 8, 18, 11, 11, 11, 12, 7, 4, 3, 6, 7, 10. Average $9·75 = pn = u.$ $e^{-u} = ·00005842.$

				$e^{-u} . u^r/r!$	Forecast.	Actual.
0	.	.	.	·00006 $\times 20$	$= ·001$	0
1 to 4	.	.	.	·0343	$= ·7$	3
5 ,, 9	.	.	.	·4564	$= 9·1$	6
10 ,, 14	.	.	.	·4408	$= 8·8$	8
15 ,, 19	.	.	.	·0683	$= 1·4$	3
20	.	.	.	small	—	0

Note added in 1936.—The result is obtained also by a hypothesis independent of p and n separately.   Write $q(t)$ for the chance that no event occurs in time or space interval t.   Then $q(t_1) \times q(t_2) = q(t_1 + t_2)$, and so on, whence $q(t) = \{q(1)\}^t = e^{-ht}$, say.   Chance of one event between t and $t + dt = -\frac{dq}{dt}dt = he^{-ht}dt$.   Write $P(n,T)$ for chance of n occurrences in T.

$P(n + 1,T) =$ chance of n in t multiplied by chance of 1 in $T - t$

$$= \int_0^T P(n,t) . he^{-h(T-t)}dt$$

$$\therefore P(1,T) = \int_0^T P(o,t)he^{-h(T-t)} dt = \int_0^T e^{-ht}he^{-h(T-t)}dt = hTe^{-hT}$$

$$\therefore P(2,T) = \int_0^T hte^{-ht}he^{-h(T-t)} dt = \tfrac{1}{2}h^2T^2e^{-ht}$$

Continuing we have $P(n,T) = \frac{1}{n!}(hT)^n e^{-hT} = \frac{u^n e^{-u}}{n!}$, where $u = hT =$ average of series.

$\therefore hT$ is average number in T, h is number per unit interval, $1/h$ is mean interval and is the only datum required.

CHAPTER III.

THE LAW OF GREAT NUMBERS (THE GENERALISED LAW OF ERROR).

So far we have treated the normal curve of error as the limit of the binomial $(q + p)^n$, and shown applications of its integral to cases where p had a definite meaning. The same equation $y = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$, however, is found as the result of much wider hypotheses, and it is the main purpose of this chapter to develop them.

•Before proceeding to the general law there are some important propositions to consider as to the relation between the standard deviation of a sum or average of magnitudes selected from a large group or groups, and the standard deviations of the magnitudes themselves. These propositions (pp. 287–9) depend only on the fundamental laws of probability, and are independent of any process of limits or of neglect of small quantities.

Standard Deviation and Mean Cube of Error of a Sum and Average.

Let $u_1, u_2 \ldots u_t \ldots u_{m_1}$ be m_1 measurements which form a frequency group, and let $\bar{u}$ be their average and σ_u their standard deviation.

Let $u_t = \bar{u} + u_t'$.

Then $\qquad m_1\bar{u} = Su_t \quad$ and $\quad \therefore Su_t' = 0,$

and $\quad m_1\sigma_u^2 = Su'_t{}^2 = S(u_t - \bar{u})^2 = Su_t^2 - 2\bar{u} . Su_t + m_1\bar{u}^2$
$$= Su_t^2 - 2\bar{u} . m_1\bar{u} + m_1\bar{u}^2$$

and $\qquad\qquad\quad$ • $\qquad \therefore Su_t^2 = m_1(\sigma_u^2 + \bar{u}^2) \quad . \quad . \quad . \quad . \quad . \quad . \quad (32)$

287

Let $v_1,\ v_2 \ldots v_t \ldots v_{m_2}$ be m_2 measurements in a second frequency curve, whose average and standard deviation are $\bar{v}$ and σ_v.

Now select at random one object from each group, say u_s and v_t. Required the average and standard deviation of the group formed by all possible values of $\overline{u_s + v_t}$, every double selection being quite independent of every other. Let H_2 be the average and s_2 the standard deviation of this group.

We will suppose that an indefinitely great number of independent selections is made, so that in the new group the $m_1 \times m_2$ possible values of $\overline{u_s + v_t}$ occur with equal frequency.

Then $H_2 \times m_1 \times m_2$

$$
\begin{aligned}
&= (u_1+v_1) + \ldots + (u_1+v_t) + \ldots + (u_1+v_{m_2}) \quad m_2 \text{ terms in} \\
&+ (u_2+v_1) + \ldots + (u_2+v_t) + \ldots + (u_2+v_{m_2}) \qquad \text{each} \\
&\qquad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \qquad \text{of } m_1 \text{ lines} \\
&\qquad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\
&+ (u_{m_1}+v_1) + \ldots + (u_{m_1}+v_t) + \ldots (u_{m_1}+v_{m_2}) \\
&= m_2 . Su_t + m_1 . Sv_s = m_2 m_1 \bar{u} + m_1 m_2 \bar{v}
\end{aligned}
$$

$$\therefore H_2 = \bar{u} + \bar{v} \quad \cdots \cdots \cdots \cdots \cdots \quad (33)$$

and $m_1 m_2 (s_2^2 + H_2^2) = S(u_s + v_t)^2$

$$
\begin{aligned}
&= (u_1+v_1)^2 + \ldots + (u_1+v_{m_2})^2 + (u_2+v_1)^2 + \ldots + (u_2+v_{m_2})^2 + \ldots \\
&= m_2 Su_t^2 + m_1 Sv_t^2 + 2 . Su_t . Sv_t \\
&= m_2 m_1 (\sigma_u^2 + \bar{u}^2) + m_1 m_2 (\sigma_v^2 + \bar{v}^2) + 2 m_1 \bar{u} . m_2 \bar{v} \\
&= m_1 m_2 \{ (\sigma_u^2 + \sigma_v^2) + (\bar{u} + \bar{v})^2 \}
\end{aligned}
$$

$$\therefore \qquad\qquad s_2^2 = \sigma_u^2 + \sigma_v^2 \quad \cdots \cdots \cdots \quad (34)$$

If the group was formed by the difference $\overline{u_s - v_t}$ instead of the sum, we should obtain in a similar way $H_2 = \bar{u} - \bar{v}$, but $s_2^2 = \sigma_u^2 + \sigma_v^2$ as before.

Next let the sum (or difference) be formed from three groups, the averages and standard deviations being $\bar{u}, \bar{v}, \bar{w}$ and $\sigma_u, \sigma_v, \sigma_w$ for the groups, and H_3, s_3 for the sum or difference.

Then

$$H_3 = \bar{u} \pm \bar{v} \pm \bar{w}, \quad \cdots \cdots \cdots \quad (35)$$

as can be readily shown.

We can obtain s_3 by supposing u_s and v_t first combined, and then a w added, and using the formula already proved twice over.

$$s_3^2 = s_2^2 + \sigma_w^2 = \sigma_u^2 + \sigma_v^2 + \sigma_w^2 \quad \cdots \cdots \quad (36)$$

and the formula can be extended by induction to any number of groups.

A very important case is when the standard deviations of the original groups are equal, so that $\sigma_u = \sigma_v = \ldots = \sigma$, say.

If the sum is formed from n such groups, and its standard deviation is s, we have

$$s^2 = s_n{}^2 = \sigma^2 + \sigma^2 + \text{to } n \text{ terms} = n\sigma^2$$

and $\therefore s = \sigma \cdot \sqrt{n}$ . . . . (37)

Next, instead of taking the sum of the n measurements, let us take their average. Every term in the composite group is then to be divided by n, and therefore the standard deviation of the group of averages, σ_a say, will be the standard deviation of the group of sums divided by n.

$$\therefore \ \sigma_a = \frac{s}{n} = \frac{\sigma}{\sqrt{n}}$$ . . . . (38)

Finally, if the average is taken of n items, all selected independently from the same (indefinitely large) initial group, so that the chance of selecting any one of the n items is not affected by previous selections, we have still $\sigma_a = \dfrac{\sigma}{\sqrt{n}}$.

In the following paragraphs it is assumed that the original measurements are all from the averages of their groups, and that therefore $0 = \bar{u} = \bar{v} = \ldots$, and $0 = Su = Sv = \ldots$

The mean cube for the sum of u_s and v_t is $\dfrac{1}{m_1 m_2} S(u_s + v_t)^3$

$$= \frac{1}{m_1 m_2}\{m_2 Su^3 + m_1 Sv^3 + 3 \cdot SvSu^2 + 3 \cdot SuSv^2\} = \frac{1}{m_1}Su^3 + \frac{1}{m_2}Sv^3$$

$= {}_u\mu_3 + {}_v\mu_3$, the sum of the third moments about the average of the groups.

Hence M_3, the third moment of the sum of n items, all from one group, $= n\mu_3$, where μ_3 is the third moment of the group, and for the sum

$$\kappa = \frac{M_3}{s^3} = \frac{n\mu_3}{n^{\frac{3}{2}}\sigma^3} = \frac{\kappa'}{\sqrt{n}}$$ . . . . . (39)

where κ' is for the group the value of " κ " as defined in formula (8).

κ is the same for the sum and for the average of n items.

Genesis of the Curve of Error.

We now proceed to the analysis which leads to the application of the curve of error. A quite simple case, which links up the two parts of this chapter, is as follows.

If the original groups are represented by normal curves of error, it can be shown that their sum and average are also normal.

For, if we write $x_t = u_t + v_t$, the chance of the concurrence of values of the parts u_t, v_t is

$$\frac{\mathrm{I}}{\sigma_u \sqrt{2\pi}} e^{-\frac{u^2_t}{2\sigma_u^2}} \times \frac{\mathrm{I}}{\sigma_v \sqrt{2\pi}} e^{-\frac{v^2_t}{2\sigma_v^2}}$$

$$= \frac{\mathrm{I}}{2\pi\sigma_u\sigma_v} e^{-\frac{1}{2}\left(\frac{u_t^2}{\sigma_u^2} + \frac{(x_t - u_t)^2}{\sigma_v^2}\right)}$$

The whole chance of $x_t(+\delta x)$ is obtainable by integrating this expression for all values of u, and equals

$$\frac{\mathrm{I}}{2\pi\sigma_u\sigma_v} \int_{-\infty}^{\infty} e^{-\frac{\sigma_u^2 + \sigma_v^2}{2\sigma_u^2\sigma_v^2}\left(u - \frac{\sigma_u^2 x_t}{\sigma_u^2 + \sigma_v^2}\right)^2} e^{-\frac{x_t^2}{2(\sigma_u^2 + \sigma_v^2)}} du . \delta x$$

$$= \left(\text{writing } u' \text{ for } u - \frac{\sigma_u^2 x_t}{\sigma_u^2 + \sigma_v^2}\right),$$

$$\frac{\mathrm{I}}{2\pi\sigma_u\sigma_v} . e^{-\frac{x_t^2}{2(\sigma_u^2 + \sigma_v^2)}} . \delta x . \int_{-\infty}^{\infty} e^{-\frac{\sigma_u^2 + \sigma_v^2}{2\sigma_u^2\sigma_v^2} u'^2} du'$$

$$= (\text{using formula p. 268}) \frac{\mathrm{I}}{s_2 \sqrt{2\pi}} e^{-\frac{x_t^2}{2s_2^2}} . \delta x, \text{ where } s_2{}^2 = \sigma_u{}^2 + \sigma_v{}^2.$$

The chance of the value x is, therefore, $\dfrac{\mathrm{I}}{s_2 \sqrt{2\pi}} e^{-\frac{x^2}{2s_2^2}}$. (40)

The process is easily generalised by induction, and the chances of obtaining x from the sum and average of n independent selections from a normal curve whose standard deviation is σ are respectively

$$\frac{\mathrm{I}}{\sigma\sqrt{2\pi n}} e^{-\frac{x^2}{2n\sigma^2}} \text{ and } \frac{\sqrt{n}}{\sigma\sqrt{2\pi}} e^{-\frac{nx^2}{2\sigma^2}} \quad . \quad . \quad . \quad (41)$$

The same result is obtainable as a first approximation when the original curves are not normal, but satisfy certain condi-

tions which are obtained in the analysis. The result is so important that two proofs are given in the following paragraphs.

Proof by the Multinomial Theorem.

In this proof it is shown that the moments obtained by an extension of the method of the preceding paragraphs (formulæ (33) to (39)) are for all orders the same as those of a normal curve of error with appropriate standard deviation.

Let there be n elemental groups containing $m_1, m_2 \ldots m_n$ measurable things respectively; in any, the tth, group, let the average, the standard deviation and the moments about the average be $\bar{u}_t, \sigma_t, {}_t\mu_2, {}_t\mu_3 \ldots$, and let the items be $\bar{u}_t + {}_tu_1, \ \bar{u}_t + {}_tu_2, \ldots, \bar{u}_t + {}_tu_s \ldots$.

Then ${}_tu_1 + {}_tu_2 + \ldots + {}_tu_s \ldots = 0$.

One item is selected at random from each group, and n such items are added; it is assumed that the selections from different groups are independent of each other, and that the chance of obtaining a particular magnitude from one group is not affected by previous selections.

In the sth selection the sum is $H + E_s$, where

$$H = \bar{u}_1 + \bar{u}_2 + \ldots + \bar{u}_n, \text{ and } E_s = {}_1u_s + {}_2u_s + \ldots + {}_nu_s.$$

Let $s, M_2, M_3 \ldots$ be the standard deviation and moments of the frequency curve of E_s, that is, of the frequency curve of the sum.

$M_2 = s^2 =$ mean of all possible values of $({}_1u_s + {}_2u_s + \ldots + {}_nu_s)^2$.

There are $m_1 \times m_2 \times \ldots \times m_n = N$, say, such values. Then generalising the process of p. 288,

$$M_2 = s^2 = \frac{1}{N}\left\{\frac{N}{m_1}S_1u^2 + \frac{N}{m_2}S_2u^2 + \ldots\right\} + \frac{2}{N}\left\{\frac{N}{m_1m_2}S_1u \cdot S_2u + \ldots\right\}$$

$$= \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2, \text{ since } 0 = S_1u = S_2u = \ldots$$

Similarly

$$M_3 = \frac{1}{N}S({}_1u_s + {}_2u_s + \ldots + {}_nu_s)^3$$

$$= \frac{1}{N}\left\{\frac{N}{m_1}S_1u^3 + \frac{N}{m_2}S_2u^3 + \ldots\right\} + \frac{3}{N}\left\{\frac{N}{m_1m_2}S_1uS_2u^2 + \ldots\right\}$$

$$= {}_1\mu_3 + {}_2\mu_3 + \ldots + {}_n\mu_3, \text{ since } 0 = S_1u, \text{ etc.} \quad \ldots \ldots \quad (42)$$

and $M_4 = \dfrac{1}{N}\left\{\dfrac{N}{m_1}\cdot S_1 u^4 + \ldots\right\} + \dfrac{4}{N}\left\{\dfrac{N}{m_1 m_2} S_1 u^3 S_2 u + \ldots\right\}$

$$+ \dfrac{12}{N}\left\{\dfrac{N}{m_1 m_2 m_3} S_1 u^2 S_2 u S_3 u + \ldots\right.$$

$$+ \dfrac{6}{N}\left\{\dfrac{N}{m_1 m_2}\cdot S_1 u^2 \cdot S_2 u^2 + \ldots\right\}$$

$$+ \dfrac{24}{N}\left\{\dfrac{N}{m_1 m_2 m_3 m_4} S_1 u S_2 u S_3 u S_4 u + \ldots\right\}$$

$$= {}_1\mu_4 + {}_2\mu_4 + \ldots + {}_n\mu_4 + 6\,(\sigma_1^2\sigma_2^2 + \sigma_1^2\sigma_3^2 + \ldots +)\ . \quad (43)$$

$\therefore\ M_4 - 3s^4 = S_t\mu_4 + 6\,(\sigma_1^2\sigma_2^2 + \ldots) - 3\,(\sigma_1^2 + \sigma_2^2 + \ldots)^2 = S({}_t\mu_4 - 3\sigma_t^4)$.

If the standard deviations and moments of the elemental curves are equal, so that $\sigma_1 = \sigma_2 = \ldots = \sigma$, ${}_1\mu_3 = {}_2\mu_3 = \ldots = \mu_3$ etc., we have

s for the sum $= \sigma\sqrt{n}$, σ_a for the average $= \dfrac{\sigma}{\sqrt{n}}$, . . (44)

κ, for sum or average, $= \dfrac{M_3}{s^3} = \dfrac{n\mu_3}{n^{\frac{3}{2}}\sigma^3} = \dfrac{\kappa'}{\sqrt{n}}$ . . . (45)

where κ' is the "κ" for the elemental curves,

$$\kappa_2 - 3 = \dfrac{M_4}{s^4} - 3 = \dfrac{n(\mu_4 - 3\sigma^4)}{n^2\sigma^4} = \dfrac{1}{n}\left(\dfrac{\mu_4}{\sigma^4} - 3\right)$$

$$= \dfrac{\text{sum of the "}\kappa_2' - 3\text{"s for elemental curves}}{n} \quad . . (46)$$

Hence κ tends to zero as $\sqrt{n}$ becomes large, and κ_2 may be taken as 3 if $\dfrac{1}{n}$ is negligible.

To find higher moments, we need to evaluate M_t for any integer t; that is the mean of $({}_1 u + {}_2 u + \ldots + {}_n u)^t$, which (by the multinomial theorem *) is the mean of

$$S\dfrac{t\,!}{n_1\,!\,n_2\,!\ldots}\cdot{}_1 u^{n_1}\cdot{}_2 u^{n_2}\ldots.$$

---

* The multinomial theorem is an extension of the binomial theorem; the following is an outline of the proof.

The product of t factors

$$(a_1 + b_1 + c_1 + \ldots)\,(a_2 + b_2 + c_2 + \ldots)\ldots(a_t + b_t + c_t + \ldots)$$

= the sum of all possible terms such as $a_1\,a_2\,b_3\,c_4\,d_5\,b_6\ldots k_t$, each suffix occurring once.

The number of such terms in which an a occurs n_1 times, a b n_2 times . . . is the number of permutations of t things taken altogether in which n_1 are alike, n_2 alike . . . i.e. $\dfrac{t\,!}{n_1\,!\,n_2\,!\ldots}$. Now write $a_1 = a_2 = \ldots = {}_1 u$, $b_1 = b_2 = \ldots = {}_2 u$, etc., and we obtain the result in the text.

when all possible terms subject to the condition $n_1 + n_2 +,\ldots = t$ are summed.

First take the case where t is even.

M_{2t} = sum of means of the terms $\dfrac{(2t)!}{n_1!\,n_2!\ldots}\,{}_1u^{n_1}\cdot{}_2u^{n_2}\ldots$

when $n_1 + n_2 + \ldots = 2t$

= sum of terms $\dfrac{(2t)!}{n_1!\,n_2!\ldots}\cdot$ mean ${}_1u^{n_1}\times$ mean ${}_2u^{n_2}\times\ldots,$

since from the independence of selection from the different elemental curves each ${}_1u$ occurs with each ${}_2u \ldots$ with equal frequency.

$\therefore M_{2t}$ = sum of terms

$$\frac{(2t!)}{n_1!\,n_2!\ldots}\cdot{}_1\mu_{n_1}\cdot{}_2\mu_{n_2}\cdots$$

Now restrict the analysis to the case, where the standard deviations and moments of the elemental curves are equal, so that ${}_1\mu_{n_1} = {}_2\mu_{n_1} = \ldots = \mu_{n_1}$, etc.

Let there be f factors ${}_1\mu_{n_1},\ {}_2\mu_{n_2}\ldots$, in any selected term. Then such a term occurs ${}_nC_f$ times in the various guises

${}_1\mu_{n_1}\times{}_2\mu_{n_2}\times{}_3\mu_{n_3}\cdots,\ {}_1\mu_{n_2}\times{}_2\mu_{n_1}\times{}_3\mu_{n_3}\cdots,\ {}_1\mu_{n_3}\times{}_2\mu_{n_2}\times{}_3\mu_{n_1}\cdots,$

each of which is identical with

$$\mu_{n_1}\times\mu_{n_2}\times\mu_{n_3}\cdots.$$

Hence M_{2t} = Sum of terms

$${}_nC_f\frac{2t!}{n_1!\,n_2!\ldots}\,\mu_{n_1}\times\mu_{n_2}\times\ldots,$$

where all values are taken subject to the condition

$$n_1 + n_2 + \ldots = 2t.$$

Now $s^2 = n\sigma^2,\ s^{2t} = n^t\cdot\sigma^{2t}$, and

$${}_nC_f = n(n-1)\ldots(n-f+1)/f!$$

$\therefore\ \dfrac{M_{2t}}{s^{2t}}$ = sum of terms

$$\frac{(2t)!}{n_1!\,n_2!\ldots}\cdot\frac{\left(1-\dfrac{1}{n}\right)\left(1-\dfrac{2}{n}\right)\ldots\left(1-\dfrac{f-1}{n}\right)}{f!}\cdot\frac{n^f}{n^t}\cdot\frac{\mu_{n_1}}{\sigma^{n_1}}\cdot\frac{\mu_{n_2}}{\sigma^{n_2}}\cdots$$

It is now necessary to restrict the elemental curves so as to satisfy the condition that $\frac{\mu_p}{\sigma^p}$ is finite for all values of p, *i.e.* that Mean $\left(\frac{u}{\sigma}\right)^p$ is finite, or that the effective range of the curve is comparable with its standard deviation. We have then to consider which of the possible terms is finite and which of order $\frac{1}{n}$ or higher.

Since μ_1 is o, each of n_1, n_2, etc. is 2 or more in every term that is not identically zero; hence since the sum of the f terms n_1, n_2 . . . is $2t$, the greatest possible number of such terms is t and $f \not> t$.

If $f < t$, the fraction $\frac{n^f}{n^t}$ is of order $\frac{1}{n}$ or higher.

If $f = t$, then $2 = n_1 = n_2 \ldots$, and we obtain, as the only term when $\frac{1}{n}$ is neglected,

$$\frac{(2t)!}{2^t} \cdot \frac{1}{f!} \cdot \frac{n^f}{n^t} \cdot \left(\frac{\mu_2}{\sigma^2}\right)^f = \frac{2t!}{2^t t!},$$

since $\mu_2 = \sigma^2$ and

$$\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) \ldots \left(1 - \frac{f-1}{n}\right)$$

is between $\qquad$ 1 and $1 - \dfrac{f(f-1)}{2n}$.

Hence $\quad M_{2t} = s^{2t} \cdot \dfrac{(2t)!}{2^t t!} = 1 \cdot 3 \cdot 5 \ldots (2t-1) s^{2t}$ . . . (47)

when terms involving $\frac{1}{n}$ are neglected.

By a similar argument

$$\frac{M_{2t+1}}{s^{2t+1}} = \text{Sum of terms } \frac{(2t+1)!}{n_1! \, n_2! \ldots} \cdot \frac{1}{f!} \cdot \frac{n^f}{n^{t+\frac{1}{2}}} \left(\frac{\mu_{n_1}}{\sigma^{n_1}}\right)\left(\frac{\mu_{n_2}}{\sigma^{n_2}}\right) \ldots$$

Here there is no term not involving a power of n in the denominator, and the greatest term is found when one of the quantities n_1, n_2 . . . $= 3$, and each of the others $= 2$; so that

$2t + 1 = n_1 + n_2 + \ldots = 2(f - 1) + 3 = 2f + 1$, and $f = t$,

and we have f equal terms obtained by putting n_1, n_2 . . . successively $= 3$

Then $\qquad \dfrac{M_{2t+1}}{s^{2t+1}} = f \times \dfrac{(2t+1)\,!}{2^{t-1}3\,!} \cdot \dfrac{1}{t!\,\sqrt{n}} \cdot \dfrac{\mu_3{}^{t-1}\cdot\mu_3}{\sigma^{2t+1}}$,

$$= \frac{t}{3}\cdot 1\cdot 3\cdot 5\ldots \overline{2t+1}\cdot \frac{\mu_3}{\sqrt{n}\,\sigma^3} \quad \ldots \quad (48)$$

$\therefore M_{2t+1}{}' = 0$ if terms involving $\dfrac{1}{\sqrt{n}}$ are neglected,

and $\quad M_{2t+1} = \dfrac{t}{3}\cdot 1\cdot 3\cdot 5\ldots \overline{2t+1}\cdot M_3\cdot s^{2t-2} \quad \ldots \ldots (49)$

since $\quad \dfrac{M_3}{s^3} = \dfrac{\mu_3}{\sqrt{n}\cdot\sigma^3}$, if terms in $\dfrac{1}{\sqrt{n}}$ are retained, and terms in $\dfrac{1}{n}$ neglected.

These moments are (see formula (23) and Appendix, Note 6) precisely those which are obtained from the curve

$$y = \frac{1}{s\sqrt{2\pi}}\cdot e^{-\frac{x^2}{2s^2}},$$

if $\dfrac{1}{\sqrt{n}}$ is neglected, and from the curve

$$y = \frac{1}{s\sqrt{2\pi}}\left[1 - \frac{\kappa}{2}\left(\frac{x}{s} - \frac{1}{3}\cdot\frac{x^3}{s^3}\right)\right]e^{-\frac{x^2}{2s^2}},$$

if $\dfrac{1}{\sqrt{n}}$ is retained and $\dfrac{1}{n}$ neglected where $\kappa = \dfrac{M_3}{s^3}$.

Hence, if we may take identity of standard deviations and of all moments as implying identity of curves, these equations are the first and second approximations to the curve of frequency required.

Professor Edgeworth's Proof.

The proof given by Professor Edgeworth (" Law of Error," *Camb. Phil. Trans.*, Vol. XX., Part I., 1904) is briefer and more ·general, but it involves rather more difficult mathematical conceptions, which it was the intention of the analysis just given (which is essentially based on Edgeworth's work) to avoid.

Edgeworth gives a formula for any number of successive approximations, but the outline which follows is confined to the first two.

With the same notation and conditions as before,

Let $E_t = {}_1u_t + {}_2u_t + \ldots n u_t$.

Let a be any fixed small quantity, only used to select terms of the same dimensions,

Then $e^{aE_t} = e^{a \cdot {}_1u_t} \cdot e^{a \cdot {}_2u_t} \cdot e^{a \cdot {}_3u_t} \ldots$ identically.

The mean value of $e^{a \cdot {}_1u_t}$, that is the mean of

$$\left(1 + a \cdot {}_1u + \frac{a^2}{2} \cdot {}_1u^2 + \frac{a^3}{3!} \cdot {}_1u^3 + \ldots\right)$$

$$= 1 + a \cdot {}_1\mu_1 + \frac{a^2}{2} \cdot {}_1\mu_2 + \frac{a^3}{3!} \cdot {}_1\mu_3 + \ldots,$$

where $\quad {}_1\mu_1 = 0.$

Since the selections from the different elemental curves are independent, the mean of the product of $e^{a \cdot {}_1u_t} \times e^{a \cdot {}_2u_t} \times \ldots =$ the product of their means.

$$\therefore \ 1 + a \cdot M_1 + \frac{a^2}{2} M_2 + \frac{a^3}{3!} M_3 + \ldots = \text{Product of } n \text{ factors}$$

such as $\qquad \left(1 + \frac{a^2}{2} \cdot {}_t\mu_2 + \frac{a^3}{3!} \cdot {}_t\mu_3 + \ldots\right)$

$$\therefore \ \log\left(1 + aM_1 + \frac{a^2}{2}M_2 + \ldots\right)$$

$$= S_{t=1}^{t=n} \log\left(1 + \frac{a^2}{2} \cdot {}_t\mu_2 + \frac{a^3}{3!} \cdot {}_t\mu_3 + \ldots\right)$$

$$= \frac{a^2}{2}S_t\mu_2 + \frac{a^3}{6} \cdot S_t\mu_3 + \frac{a^4}{24}S_t\mu_4 + \ldots - \frac{1}{2}S\left(\frac{a^2}{2} \cdot {}_t\mu_2 + \ldots\right)^2$$

$$\therefore \ 1 + aM_1 + \frac{a^2}{2}M_2 + \ldots$$

$$= e^{\frac{a^2}{2} \cdot S_t\mu_2} \cdot e^{\frac{a^3}{6}S_t\mu_3} \cdot e^{\frac{a^4}{24}(S_t\mu_4 - 3S({}_t\mu_2)^2)} \ldots$$

$$= \left(1 + \frac{a^2}{2}S_t\mu_2 + \ldots + \frac{1}{p!}\left(\frac{a^2}{2}S_t\mu_2\right)^p + \ldots\right) \cdot \left(1 + \frac{a^3}{6}S_t\mu_3 + \ldots\right) \cdot$$

$$\left(1 + \frac{a^4}{24}\left\{S_t\mu_4 - 3S({}_t\mu_2)^2\right\} + \ldots\right)\ldots$$

Equate coefficients up to a^4.

$$M_1 = 0$$

$$s^2 = M_2 = S_t\mu_2 = S\sigma_t^2 = n\sigma^2, \text{ if } \sigma^2 \text{ is the mean of } \sigma_1{}^2, \sigma_2{}^2 \ldots$$

$$M_3 = S_t\mu_3 = n\mu_3, \text{ if } \mu_3 \text{ is mean of } {}_1\mu_3, {}_2\mu_3 \ldots$$

$$\frac{M_4}{24} = \frac{1}{2} \cdot \left(\frac{1}{2} \cdot S_t\mu_2\right)^2 + \frac{1}{24}\left(S_t\mu_4 - 3S({}_t\mu_2)^2\right).$$

$\therefore M_4 - 3s^4 = S({}_t\mu_4 - 3\sigma_t{}^4)$, and $\kappa_2 - 3 = \dfrac{M_4}{s^4} - 3 = \dfrac{1}{n^2\sigma^4}S\left(\dfrac{{}_t\mu_4}{\sigma_t{}^2} - 3\right)\sigma_t{}^4$

$= \dfrac{1}{n}(\kappa_2' - 3)$, if $\kappa_2' - 3$ is mean $\left(\dfrac{{}_t\mu_4}{\sigma_t{}^4} - 3\right)\left(\dfrac{\sigma_t}{\sigma}\right)^4$.

$\kappa = \dfrac{M_3}{s^3} = \dfrac{n\mu_3}{n^{\frac{3}{2}}\sigma^3} = \dfrac{\kappa'}{\sqrt{n}}$, where $\kappa' = \dfrac{\mu_3}{\sigma^3}$.

Hence

$1 + \dfrac{a^2}{2}s^2 + \dfrac{a^3}{3!}M_3 + \ldots + \dfrac{a^t}{t!}M_t \ldots$

$= e^{\frac{1}{2}a^2 s^2}\left(1 + \dfrac{1}{6}a^3 \cdot s^3 \cdot \dfrac{1}{\sqrt{n}}\kappa' + \ldots\right)\left(1 + \dfrac{1}{24}a^4 s^4 \cdot \dfrac{1}{n}(\kappa_2' - 3) + \ldots\right)$

On the right-hand side of the equation in every case the index of a equals the suffix of μ or the sums of the suffixes of powers or products of μ's.

Now assume that throughout the elemental curves $\dfrac{\mu_p}{\sigma^p}$ is finite for all values of p, and it results that the coefficient of $a^p \cdot s^p$ contains the factor $\dfrac{1}{n^{\frac{1}{2}(p-2)}}$, as has been worked out above up to a^4.

Neglect $\dfrac{1}{n^{\frac{1}{2}}}$ and all higher powers,

$1 + \dfrac{a^2}{2}s^2 + \ldots + \dfrac{a^t}{t!}M_t + \ldots = 1 + \dfrac{1}{2}a^2 s^2 + \ldots + \dfrac{1}{t!}a^{2t} \cdot \dfrac{s^{2t}}{2^t} + \ldots$

$\therefore$ every odd moment, $M_{2t+1}, = 0$

and an even moment, $M_{2t,} = (2t)!\dfrac{s^{2t}}{t!\,2^t} = 1.3\ldots(2t-1).s^{2t}$ (50)

as in the normal curve of error (formula (23)).

Now retain $\dfrac{1}{\sqrt{n}}$ and neglect $\dfrac{1}{n}$.

M_{2t} is as before.

$\dfrac{M_{2t+1}}{(2t+1)!} = \dfrac{1}{(t-1)!\,2^{t-1}}s^{2t-2} \cdot \dfrac{1}{6} \cdot s^3 \cdot \dfrac{\kappa'}{\sqrt{n}},$

$M_{2t+1} = \dfrac{(2t+1)!}{(t-1)!\,2^{t-1}} \cdot \dfrac{s^{2t-2}}{6} \cdot M_3 = \dfrac{t}{3} \cdot 1.3.5\ldots(2t+1) \cdot M_3 \cdot s^{2t-2},$

that is, the $(2t+1)^{\text{th}}$ moment of the curve

$$\dfrac{1}{s\sqrt{2\pi}}\left\{1 - \dfrac{\kappa}{2}\left(\dfrac{x}{s} - \dfrac{1}{3}\dfrac{x^3}{s^3}\right)\right\}e^{-\frac{x^2}{2s^2}} \quad . \quad . \quad . \quad . \quad (51)$$

(see Appendix, Note 6).

Hence, by the test of equality of moments, the curve of frequency of the sum or average of n selections under the given conditions has for its first approximation the normal curve, when $\dfrac{1}{\sqrt{n}}$ is neglected, and for its second approximation the skew curve already given.

Further approximations, which so far have been found mainly of theoretic interest only, are given by Edgeworth.

Statement of the Generalised Law of Error, or the Law of Great Numbers.

The theorems now proved can be summarised as follows, the conditions of validity being restated and amplified.

Let there be a large number (n) of elemental groups, each of which can be represented by a frequency locus, such that the chance of obtaining a magnitude U by selection from a group is a function of U.

Form a total, H, of n things, one selected from each group, so that the selection from one group has no (or very slight) effect on the selection from another ; and obtain many values of H by repeating the process, in such a way that the selections which make one value of H are not affected by the selections which make other values.*

Then if the frequency loci of the elemental groups satisfy certain conditions, the frequency locus of H has a definite form to which $y = \dfrac{1}{s\sqrt{\pi 2}}\, e^{-\frac{x^2}{2s^2}}$ is a first, and

$$y = \frac{1}{s\sqrt{2\pi}}\, e^{-\frac{x^2}{2s^2}}\left[1 - \frac{\kappa}{2}\left(\frac{x}{s} - \frac{1}{3}\frac{x^3}{s^3}\right)\right]$$

is a second approximation, where s^2 is the second moment and κs^3 the third moment of the locus.

The frequency locus of the average of the n magnitudes is of the same form as that of the sum, and κ has the same value in both cases.   If s_a is the standard deviation of the average,

---

* If the selections from one elemental group are not independent, but the magnitudes tend to come in batches, then more values of H are necessary to obtain any given approximation to its final frequency form when an indefinitely large number of values are taken.

$s_a = \dfrac{s}{n}$, where $s = \sigma \sqrt{n}$ and $s_a = \dfrac{\sigma}{\sqrt{n}}$ if σ is typical of the standard deviation of the elemental curves. κ is of the order $\dfrac{1}{\sqrt{n}}$ in comparison with 1 in the frequency equation of H, and only the first approximation is necessary when n is very great or when the elemental curves are symmetrical, in which case $\kappa = 0$.

The condition that must be satisfied by the elemental curves is that, if μ_p is the pth moment and σ the standard deviation of any one of them, $\dfrac{\mu_p}{\sigma^p}$ is a small finite number *$\left(\text{that can be neglected when multiplied by} \dfrac{1}{n^{\frac{p}{2}-1}} \right)$ for all values of p;

this is secured when the great bulk of the frequency curve is on a base containing only a small multiple (1, 2 or 3) of its standard deviation to left and right of its average. This condition is quite generally satisfied by ordinary frequency groups when n is at all large.

The first and the second approximations are only valid for moderate values of $\dfrac{x}{s}$, since beyond these the contributions of further approximations become sensible; it is only the central portion of the frequency curve of H so generated that is determinable; the outer portions have no general form, and it can only be postulated that their aggregate volume is small, and that the chance of exceeding, say $3s$, is negligible. The range that is to be understood by "the central portion" depends on the value of n; as the number of independent elements increases, so the range of the determinable form extends. In ordinary cases with n as great as 100 it may perhaps be said that the frequency curve is known over a range of $2s$ on either side of the origin.

It follows that the applicability of the law of error to given observations should not be denied on the ground that the positions of extreme values do not conform to the law.

---

* More exactly it is only the difference between this ratio and the corresponding ratio in a normal curve that is involved.

Case when the Universe is Limited.

On pp. 287 *seq.* it was assumed that the selection of one item did not affect the chance of further selections.

As on p. 282, we will now examine the case where the universe from which the selection is made is limited, so far as the determination of the average is concerned.

Let a group of n things be selected at random from a group of N things, whose measurements are $a + u_1$, $a + u_2$ ..., $a + u_N$, where a is the average and $S_1^N u_t = 0$. Write $H + E$ for the sum of the measurements of the n selected things, where $H = na$.

There are $_NC_n$ equally probable values of E, such as

$$u_1 + u_2 + u_3 + \ldots + u_n$$
$$u_1 + u_3 + u_4 + \ldots + u_N$$
$$\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot$$

The sum of the values is easily seen to be zero, and therefore the mean value of $E = 0$.

Let s be standard deviation of E.

Then $_NC_n \cdot s^2 = $ sum of $_NC_n$ squares, such as $(u_1 + u_2 + \ldots + u_n)^2$ each containing n terms.

In the sum each square, such as u_t^2, occurs $\dfrac{n}{N} \times {}_NC_n$ times, and each product $2u_su_t$ occurs $\dfrac{1}{_NC_2} \times \dfrac{n(n-1)}{2} \times {}_NC_n$ times, since in all there are $n \times {}_NC_n$ squares and $\dfrac{n(n-1)}{2} \times {}_NC_n$ products.

$$\therefore \quad {}_NC_n \cdot s^2 = \frac{n}{N} \times {}_NC_n \cdot S_1^N u_t^2 + 2 \cdot \frac{n(n-1)}{N(N-1)} \cdot {}_NC_n \cdot S u_s u_t$$

$$s^2 = \frac{n}{N} \cdot N\sigma^2 + \frac{n(n-1)}{N(N-1)} \{ (S u_t)^2 - S u_t^2 \},$$

where σ is standard deviation of the universe from which selection was made,

$$= n\sigma^2 - \frac{n(n-1)}{N-1} \sigma^2, \quad \text{since } S u_t = 0$$

$$= \sigma^2 \cdot n \frac{N-n}{N-1} = \sigma^2 n \cdot \left(1 - \frac{n}{N}\right), \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad (52)$$

if $\dfrac{1}{N}$ is negligible.

Let σ_a be the standard deviation of the average $\left(\dfrac{H}{n} + \dfrac{E}{n}\right)$ of the n selections.

Then $\qquad \sigma_a = \dfrac{s}{n} = \dfrac{\sigma}{\sqrt{n}} \cdot \sqrt{\left(1 - \dfrac{n}{N}\right)}.$ . . . . . . (53)

If N is indefinitely great, we obtain $\dfrac{\sigma}{\sqrt{n}}$ as before, formula (38).

By neglecting $\dfrac{n}{N}$ we exaggerate the standard deviation.

It can be shown (Isserlis, *Stat. Journal*, 1918, pp. 75 *seq.*) that the frequency of the sum or average is very approximately normal, when N is large as is generally the case in practice.

If we use the table on p. 271 with the value $\dfrac{\sigma}{\sqrt{n}}$, we exaggerate slightly throughout the chance that a deviation exceeds any given amount.

Note.—That the law of great numbers is obtainable from the limit of the terms of $(p + q)^n$, as shown above, can be proved as a special case of the general analysis.

Let each elemental group contain qm zeros and pm units, where $p + q = 1$.

The constants of such a group are

$$\bar{x} = \frac{qm \times 0 + pm \times 1}{qm + pm} = p.$$
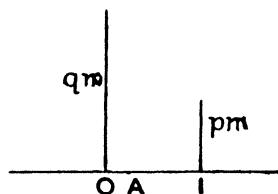
pm, qm are at distances $+ q$ and $- p$ from the average, A.

$$\mu_2 = \frac{qm\,(- p)^2 + pm\,(q)^2}{(p + q)\,m} = pq, \quad \sigma = \sqrt{pq},$$

$$\kappa' = \frac{\mu_3}{\sigma^3} = \frac{q\,(- p^3) + p\,(q)^3}{(pq)^{\frac{3}{2}}} = \frac{q - p}{\sqrt{pq}}.$$

Form a total by adding selections one from each of n such curves; this satisfies the conditions for the formation of H above.

The total has a frequency curve with average pn, standard deviation $\sigma\sqrt{n} = \sqrt{pqn}$, and $\kappa = \dfrac{\kappa'}{\sqrt{n}} = \dfrac{q - p}{\sqrt{pqn}}$, as already found, p. 264.

Illustrative Examples.

In its integral form the law of great numbers, so far as the second approximation, is

$$P\left(\pm x\right) = \frac{1}{\sigma\sqrt{2\pi}}\int_0^x e^{-\frac{x^2}{2\sigma^2}}\cdot dx \mp \frac{\kappa}{6\sqrt{2\pi}}\left\{1 - \left(1 - \frac{x^2}{\sigma^2}\right)e^{-\frac{x^2}{2\sigma^2}}\right\} *$$

$$= \frac{1}{\sqrt{2\pi}}\int_0^z e^{-\frac{1}{2}z^2}dz \mp \frac{\kappa}{6\sqrt{2\pi}}\left\{1 - (1 - z^2)\,e^{-\frac{1}{2}z^2}\right\}$$

$$= F(z) \mp \kappa f(z), \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (54)$$

where $P(x)$ is the chance of a positive deviation from the average not exceeding x, $z = \dfrac{x}{\sigma}$, $F(z)$ is tabulated on p. 271, and

$f(z) = \dfrac{1}{6\sqrt{2\pi}}\{1 - (1 - z^2)\,e^{-\frac{1}{2}z^2}\}$ is tabulated on the next page.

$\kappa = \dfrac{\mu_3}{\sigma^3}$, and μ_3 and σ are the third moment and standard deviation respectively of the curve, calculated either *a priori* or from the observations.

Eight examples follow to illustrate the method of fitting the curve to observations. In the first two (words and bricks) the genesis of the measurements leads one to expect agreement with the law of great numbers ; in the next two (skulls and plaice) application to biometrical measurements is shown ; in the next (ages) there is an indirect relation to mental phenomena ; in the last three (speeds, food consumption, and prices) the nature of the variation is complex and sporadic, and the form of the frequency curve could not be forecast.

Only the first example is worked in full.

---

* See Appendix, Note 6, for the integration.

TABLE OF VALUES OF $f(x) = \dfrac{1}{6\sqrt{2\pi}} \left\{ 1 - (1 - x^2)e^{-\frac{1}{2}x^2} \right\}$.

x	$f(x)$	x	$f(x)$	x	$f(x)$	x	$f(x)$	x	$f(x)$
·00	·0000	·50	·0225	1·00	·0665	1·50	·0935	2·00	·0935
·01	·0000	·51	·0233	1·01	·0673	1·51	·0937	2·02	·0931
·02	·0000	·52	·0241	1·02	·0681	1·52	·0939	2·04	·0927
·03	·0001	·53	·0249	1·03	·0689	1·53	·0942	2·06	·0923
·04	·0002	·54	·0258	1·04	·0697	1·54	·0944	2·08	·0919
·05	·0003	·55	·0266	1·05	·0704	1·55	·0945	2·10	·0915
·06	·0004	·56	·0275	1·06	·0712	1·56	·0947	2·12	·0911
·07	·0005	·57	·0283	1·07	·0719	1·57	·0949	2·14	·0906
·08	·0006	·58	·0292	1·08	·0727	1·58	·0951	2·16	·0902
·09	·0008	·59	·0301	1·09	·0734	1·59	·0952	2·18	·0897
·10	·0010	·60	·0310	1·10	·0741	1·60	·0953	2·20	·0892
·11	·0012	·61	·0318	1·11	·0748	1·61	·0955	2·22	·0887
·12	·0014	·62	·0327	1·12	·0755	1·62	·0956	2·24	·0882
·13	·0017	·63	·0336	1·13	·0762	1·63	·0957	2·26	·0877
·14	·0019	·64	·0345	1·14	·0769	1·64	·0958	2·28	·0872
·15	·0022	·65	·0354	1·15	·0776	1·65	·0959	2·30	·0867
·16	·0025	·66	·0363	1·16	·0782	1·66	·0959	2·32	·0862
·17	·0028	·67	·0372	1·17	·0789	1·67	·0960	2·34	·0857
·18	·0032	·68	·0381	1·18	·0795	1·68	·0960	2·36	·0853
·19	·0035	·69	·0390	1·19	·0801	1·69	·0961	2·38	·0848
·20	·0039	·70	·0399	1·20	·0807	1·70	·0961	2·40	·0843
·21	·0043	·71	·0409	1·21	·0813	1·71	·0961	2·42	·0838
·22	·0047	·72	·0418	1·22	·0819	1·72	·0961	2·44	·0833
·23	·0052	·73	·0427	1·23	·0825	1·73	·0962	2·46	·0828
·24	·0056	·74	·0436	1·24	·0831	1·74	·0962	2·48	·0823
·25	·0061	·75	·0445	1·25	·0836	1·75	·0962	2·50	·0818
·26	·0066	·76	·0455	1·26	·0842	1·76	·0961	2·52	·0814
·27	·0071	·77	·0464	1·27	·0847	1·77	·0961	2·54	·0809
·28	·0076	·78	·0473	1·28	·0852	1·78	·0961	2·56	·0804
·29	·0081	·79	·0482	1·29	·0857	1·79	·0960	2·58	·0800
·30	·0086	·80	·0491	1·30	·0862	1·80	·0960	2·60	·0795
·31	·0092	·81	·0500	1·31	·0867	1·81	·0959	2·62	·0791
·32	·0098	·82	·0509	1·32	·0871	1·82	·0958	2·64	·0787
·33	·0104	·83	·0518	1·33	·0876	1·83	·0958	2·66	·0782
·34	·0110	·84	·0527	1·34	·0880	1·84	·0957	2·68	·0778
·35	·0116	·85	·0536	1·35	·0885	1·85	·0956	2·70	·0774
·36	·0122	·86	·0545	1·36	·0889	1·86	·0955	2·72	·0770
·37	·0129	·87	·0554	1·37	·0893	1·87	·0954	2·74	·0766
·38	·0136	·88	·0563	1·38	·0897	1·88	·0953	2·76	·0762
·39	·0142	·89	·0572	1·39	·0901	1·89	·0952	2·78	·0759
·40	·0149	·90	·0581	1·40	·0904	1·90	·0950	2·80	·0755
·41	·0156	·91	·0589	1·41	·0908	1·91	·0949	2·82	·0752
·42	·0164	·92	·0598	1·42	·0912	1·92	·0948	2·84	·0748
·43	·0171	·93	·0607	1·43	·0915	1·93	·0946	2·86	·0745
·44	·0178	·94	·0616	1·44	·0918	1·94	·0945	2·88	·0742
·45	·0186	·95	·0624	1·45	·0922	1·95	·0943	2·90	·0738
·46	·0193	·96	·0632	1·46	·0924	1·96	·0942	2·92	·0735
·47	·0201	·97	·0640	1·47	·0927	1·97	·0940	2·94	·0732
·48	·0209	·98	·0649	1·48	·0930	1·98	·0938	2·96	·0730
·49	·0217	·99	·0657	1·49	·0932	1·99	·0937	2·98	·0727

x	$f(x)$	x	$f(x)$
3·00	·0724	3·80	·0671
3·20	·0702	4·00	·0668
3·40	·0687	4·20	·0666
3·60	·0677	∞	·0665

1. A lengthy book was selected, and a number of letters in each of the first completed words in 10,000 consecutive lines were noted (A); also the total number of letters in the 1000 batches obtained by adding the first 10 entries, the second 10, etc. (B) ; and 100 totals were similarly obtained by adding batches of 100 (C).

The curve of frequency of A is purely observational, and its form cannot be foretold ; that of B tends to satisfy the conditions under which the law of great numbers appears, but " n " is only 10, and unless A is nearly normal the form can only be foretold in the central region ; in C, with " n " 100, the second approximation should fit over a considerable region, and the first approximation will be sufficient if A is fairly symmetrical.

A.—DISTRIBUTION OF 10,000 WORDS ACCORDING TO THE NUMBERS OF LETTERS IN THEM.

Number of letters.	z	Observ-ations. y	xy	x^2y	x^3y	z	$F(z)$*	Diff. × 10,000
1 or ·5 to 1·5	−7	127	− 889	6,223	− 43561	−1·62	0·447	490
2 „ 1·5 „ 2·5	−6	1,792	− 10752	64,512	−387072	−1·27	·398	770
3 „ 2·5 „ 3·5	−5	1,984	− 9920	49,600	−248000	− ·92	·321	1,020
4 „ 3·5 „ 4·5	−4	1,240	− 4960	19,840	− 79360	− ·58	·219	1,280
5 „ 4·5 „ 5·5	−3	968	− 2904	8,712	− 26136	− ·23	·091⎞	1,390
6 „ 5·5 „ 6·5	−2	812	− 1624	3,248	− 6496	+ ·12	·048⎠	1,330
7 „ 6·5 „ 7·5	−1	893	− 893	893	− 893	+ ·47	·181	1,130
8 „ 7·5 „ 8·5	0	634	0	0	0	+ ·82	·294	850
9 „ 8·5 „ 9·5	1	602	+ 602	602	+ 602	+1·17	·379	570
10 „ 9·5 „ 10·5	2	460	+ 920	1,840	+ 3680	+1·52	·436	330
11 „ 10·5 „ 11·5	3	260	+ 780	2,340	+ 7020	+1·87	·469	180
12 „ 11·5 „ 12·5	4	116	+ 464	1,856	+ 7424	+2·22	·487	80
13 „ 12·5 „ 13·5	5	69	+ 345	1,725	+ 8625	+2·57	·495	30
14 „ 13·5 „ 14·5	6	21	+ 126	756	+ 4536	+2·92	·498	10
15 „ 14·5 „ 15·5	7	18	+ 126	882	+ 6174	+3·27	·499	10
16 „ 15·5 „ 16·5	8	4	+ 32	256	+ 2048	+3·62	·500	0
		10,000	−31942 +3395	163,285	−791518 +40109			
			−28547		−751409			

$\bar{x} = -2\cdot 8547$. Average is $8 - \bar{x} = 5\cdot 1453$.
$\mu_2 = 16\cdot 3285 - \bar{x}^2 = 8\cdot 1792$, $\sigma = 2\cdot 860$.
$\mu_3 = -75\cdot 1409 - 3(-2\cdot 8547)(16\cdot 3285) + 2 \ -2\cdot 8547)^3 = 18\cdot 1704$.
$k = \dfrac{\mu_3}{\mu_2^{\frac{3}{2}}} = \cdot 78$.

For calculating the moments an arbitrary origin has been taken at 8.

In fitting the normal curve $z = \dfrac{x - \bar{x}}{\sigma}$. The first entry $F(z) = \cdot 447$ shows the proportion included between the average and ·5 letters ($x = -7.5$). The normal curve gives 530 instances

---

* Table, p. 271.

below ·5 letters, and in other respects the last column is not a close approximation to y, the observations. The original curve is not normal, but it is unimodal and continuous, and in spite of its skewness the great bulk is contained in the limits $\bar{x} \pm 2\sigma$. Hence we have all the conditions for obtaining the law of great numbers if we add elements taken at random from the curve.

B.—DISTRIBUTION OF 1000 SUMS OF THE LETTERS IN 10 WORDS.

Number of letters.	z	$F(z)$	Differences × 1000.	Observations.	$f(z)$†	$F(z)$ $\mp \kappa f(z)$*	Differences × 1000.
26·5	− 2·650	·496	4	0	·078	·528	0
31·5	− 2·119	·483	13	8	·091	·520	8
36·5	− 1·588	·444	39	38	·095	·483	37
41·5	− 1·057	·355	89	97	·071	·384	99
46·5	− ·526	·201⎫	154	155	·025	·211⎫	173
51·5	+ ·005	·002⎭	203	227	·000	·002⎭	213
5?·5	+ ·536	·204	202	202	·026	·193	191
61·5	+ 1·067	·357	153	134	·072	·328	135
66·5	+ 1·598	·445	88	76	·095	·406	78
71·5	+ 2·129	·483	38	37	·091	·446	40
76·5	+ 2·660	·496	13	13	·078	·464	18
81·5	+ 3·191	·499	3	9	·069	·471	7
86·5	+ 3·722	·500	1	3	·067	·473	2
			0	1			0

For the 1000 sums the average is 51·453, $\sigma = 9\cdot4155$, $\kappa = \cdot4093$. The sums are all between 26 and 87. The calculation of the columns z, $F(z)$ and Differences are on the same method as for A. The normal curve now fits much better, and in the range 31·5 to 76·5, that is, average $\pm 2\sigma$, there is nothing to be desired ; but the formula gives too many below 31·5 and too few above 76·5.

The second approximation gives a very close fit throughout, except that it fails to stretch so as to include the one entry above 86·5 (see p. 432 for test of fit).

---

* + for negative values of z.   † Table, p. 303

X*

We should have expected to find that the standard deviation and κ of these observations were the standard deviation (2·860) and κ (·78) multiplied respectively by $\sqrt{10}$ and $\dfrac{1}{\sqrt{10}}$ (formulæ (37) and (39)).

But 2·860 × $\sqrt{10}$ = 9·04 and ·81 ÷ $\sqrt{10}$ = ·25, whereas we get from the B observations 9·42 and ·41. This points to a failure of complete independence in the aggregation of the 10 words; and analysis shows that the author's style changes from the earlier to the later part of the book, so that there is some correlation between 10 words taken consecutively. In fact, when we sum 100 words consecutively as in C, we get $\sigma = 33\cdot311$ instead of 2·86 × $\sqrt{100}$, while when the order of summation was re-arranged so as to include entries from all parts of the book in each 100, σ was 28·87, which accords with theory.

C.—DISTRIBUTION OF 100 TOTALS OF THE LETTERS IN 100 WORDS.

Number of letters.	z	F(z)	Differences × 100.	Observations.
415	−3·001	·499		
			·7	1
435	−2·400	·492		
			2·8	2
455	−1·800	·464		
			7·9	7
475	−1·200	·385		
			16·0	19
495	−·599	·225		
			22·5	25
515	−·001	·000		
			22·6	18
535	+ ·602	·226		
			15·9	18
555	+1·202	·385		
			7·9	6
575	+1·803	·464		
			2·8	3
595	+2·403	·492		
			·7	0
615	+3·003	·499		
			·1	1
635	+3·604	·500		

The agreement between formula and observations in this table is very close (see p. 432), and cannot be improved perceptibly by using the second approximation.

This experiment, which was devised with the definite

intention of illustrating the law of great numbers (and the correlation surface, formula (102)), has thus proved to be completely satisfactory, even in that it also illustrates the difficulty of securing random selection.

2. In a garden the paths are bordered by bricks originally laid (but not mortared) lengthways touching each other. After they had been exposed for some time to the influences of weather and of gardening operations, the lengths occupied by 143 sequences of 4 bricks were measured as nearly as possible to the nearest sixteenth of an inch. The causes of variation were—inequalities of the bricks as they came from the mould, inequalities in the slight interval between one and the next, displacement since they were laid, and difficulties of measurement. These causes are multiple and independent and each of small effect. It might be expected that their effects can be expressed as the sum of errors, and that the distribution of the measurements would be approximately normal and symmetrical.

DISTRIBUTION OF LENGTHS OF FOUR BRICKS.

Length.	Number of observations.	Calculated by formula (normal curve).
35	1	·7
$35\tfrac{1}{16}$	1	1·4
$35\tfrac{1}{8}$	3	2·7
$35\tfrac{3}{16}$	7	5·1
$35\tfrac{1}{4}$	11	8·0
$35\tfrac{5}{16}$	4	11·6
$35\tfrac{3}{8}$	21	15·0
$35\tfrac{7}{16}$	7	17·7
$35\tfrac{1}{2}$	30	18·3
$35\tfrac{9}{16}$	16	17·5
$35\tfrac{5}{8}$	13	14·9
$35\tfrac{11}{16}$	6	11·4
$35\tfrac{3}{4}$	11	7·9
$35\tfrac{13}{16}$	7	5·0
$35\tfrac{7}{8}$	4	2·7
$35\tfrac{15}{16}$	1	1·4
36	0	·6
	143	142·9

x^2 2

Except for an obvious tendency to give the measurements to the nearest ⅛th of an inch instead of the ₁/₁₆th, the fit is fairly satisfactory. When this tendency is corrected the fit is very good.

3. I am indebted to Professor C. G. Seligman's *Some Aspects of the Hamitic Problem in the Anglo-Egyptian Sudan* for the following measurements, whose frequency groups I analysed at his request.

SKULL AND STATURE MEASUREMENTS OF THE DINKA RACE.

Grades from average.	F(s).	Cephalic Index.		Nasal Index.		Stature.	
		Difference ×148.	Observations.	Difference ×85.	Observations.	Difference ×116.	Observations.
		·2	2*	·1	0	·1	0
Over 3σ .	·4986						
		·9	1	·5	1	·7	2
⅓σ– . .	·4938						
		2·3	2	1·3	0	1·8	1
2σ– . .	·4772						
		6·5	4	3·7	4	5·1	1
⅓σ– . .	·4332						
		13·6	14	7·8	6	10·7	6
σ– . .	·3413						
		22·2	18	12·8	13	17·4	22
σ/2 . .	·1915						
		28·3	30	16·3	27	22·2	24
0– . .	0						
		28·3	30	16·3	12	22·2	25
–σ/2 . .	·1915						
		22·2	25	12·8	8	17·4	23
–σ . .	·3413						
		13·6	13	7·8	6	10·7	6
–⅓σ . .	·4332						
		6·5	7	3·7	4	5·1	4
–2σ . .	·4772						
		2·3	2	1·3	3	1·8	1
–⅓σ . .	·4938						
		·9	0	·5	1	·7	1
Under –3σ.	·4986						
		·2	0	·1	0	·1	0
Total .		148	148	85	85	116	116
Average .		—	72·7	—	91·6	—	178·6 cm.
Standard deviation .			3·70		13·0		9·66

Except for the two extreme cases marked * the range is normal, and the deviations from the normal curve are not greater than is to be expected with so few examples.

4. The lengths of 554 plaice measured in the North Sea Fisheries Investigation gave the following results :—

Length cm.	s	F(s)	Difference × 554.	Observations.
			1·3	0
35·5	2·825	·4976		
			9·2	6
34·5	2·076	·4810		
			40·6	50
33·5	1·327	·4077		
			104·9	105
32·5	·578	·2183		
			158·6	166
31·5	— ·171	·0679		
			140·3	145
30·5	— ·920	·3212		
			72·7	61
29·5	—1·669	·4524		
			22·0	10
28·5	—2·418	·4922		
			3·9	7
27·5	—3·167	·4992		
			·4	3
26·5	—3·916	·5		
			0	1
25·5	—4·665	·5		554

Average 31·778 ; $\sigma = 1·335$.

The agreement is not close at the extremities.

•5. The number of school children of various ages in the sixth grade are given in a report of the public schools of St. Louis, U.S.A.

The following table compares the data with the first and second approximations to the law of great numbers :—

Ages.	Number of children.	Numbers calculated from 1st Approx.	2nd Approx.
10–	26	39	27
11–	201	207	204
12–	673	630	670
13–	1,001	983	995
14–	739	785	746
15–	310	323	307
16–	80	67	79
17–	13	9	15
18–	1	0	0

Average age, 13·665 ; $\sigma = 1·100$; $\kappa = ·2059$.

The first approximation fits well within 2σ of the average.

The second approximation is remarkably close to the observations (see diagram in Appendix, Note 6).

6. The speeds of 100 pedestrians were calculated from observing the time they took between two marks (*Die Schwankungen der landwirtschaftlichen Reinerträge*—Mitscherlich).

Average velocity, 1·5846 metres per second. $\sigma = ·2179$ m.

Speed.		Numbers at various speeds.	
		Calculated.	Actual.
Average +·50m or more. . . .		1·1	2
+·40 to ·50 . . .		2·3	2
+·30 ,, ·40 . . .		5·0	4
+·20 ,, ·30 . . .		9·6	11
+·10 ,, ·20 . . .		14·3	10
0 ,, ·10 . . .		17·7	18
−·10 ,, 0 . . .		17·7	20
−·20 ,, −·10 . . .		14·3	15
−·30 ,, −·20 . . .		9·6	8
−·40 ,, −·30 . . .		5·0	7
−·50 ,, −·40 . . .		2·3	3
−·50 or less		1·1	0

7. From material collected by the Working Class Cost of Living Committee, 1918, the expenditure on food in one week by 970 urban families was determined, and the results divided by the number of " equivalent adults " (where a child is taken as a fraction of an adult). The average was 10·75s ; $\sigma = 3·156$, $\kappa = ·84$.

Weekly expenditure per " unit " on food.		Actual.	Number of Families.	
			Calculated by 2nd approx.	Calculated from Pearson's Type III.*
Not exceeding 5·5s . . .		18	22	7
5·5s		107	123	122
7·5		255	233	252
9·5		245	248	250
11·5		173	168	172
13·5		101	89	95
15·5		38	51	45
17·5		17⎫	22⎫	19⎫
19·5		9⎬33	11⎬35	7⎬27
Over 21·5		7⎭	1⎭	1⎭

$\beta_1 = ·708.$ $\beta_2 = 4·035.$

8. The price of flour was determined in U.S.A. in 272 places. In five towns the price was given as 4 cents per lb., and these were evidently exceptional and are excluded. For the remaining 267 the average was 2·629 cents per lb., and $\sigma = ·3334$.

* See p. 345.

TOWNS CLASSIFIED ACCORDING TO THE PRICE OF FLOUR.

Average.					Calculated 1st Approx.	Actual.
$+3\sigma$ or more	.	.	.	.	·4	2
$+\frac{1}{2}\sigma$ to 3σ	.	.	.	.	1·3	3
$+2\sigma$,, $\frac{1}{2}\sigma$	.	.	.	.	4·4	1
$+\frac{1}{3}\sigma$,, 2σ	.	.	.	.	11·8	9
$+\sigma$,, $\frac{1}{3}\sigma$	.	.	.	.	24·5	16
$+\frac{1}{2}\sigma$,, σ	.	.	.	.	40·0	43
0 ,, $\frac{1}{2}\sigma$	.	.	.	.	51·1	67
$-\dfrac{\sigma}{2}$,, 0	.	.	.	.	51·1	47
$-\sigma$,, $-\dfrac{\sigma}{2}$	.	.	.	.	40·0	37
$\div\frac{1}{2}\sigma$,, $-\sigma$	.	.	.	.	24·5	25
-2σ ,, $-\frac{1}{2}\sigma$	.	.	.	.	11·8	9
$-\frac{1}{3}\sigma$,, -2σ	.	.	.	.	4·4	4
Less than $-\frac{1}{2}\sigma$	.	.	.	.	1·7	4

In the range average $\pm\ 2\sigma$ the agreement is fairly satisfactory and satisfies the test explained below (Chapter X).

CHAPTER IV.

APPLICATIONS OF THE LAW OF ERROR.

Precision of Sums and Averages.

It follows from the previous chapter that if n measurable things are selected at random from a universe where the sizes are distributed in a frequency group which is fairly continuous, and little of it far from its average as compared with its standard deviation (σ), then the average belongs to a frequency curve whose standard deviation is $\sigma/\sqrt{n}$ and its form approximately normal.

σ has generally to be determined from the observations themselves, and may differ from that of the universe, but only by a quantity of order $\dfrac{\sigma}{\sqrt{n}} \times \dfrac{1}{\sqrt{n}}$ (see p. 417 below).

The first illustration that follows (persons per tenement) gives 12 cases where the averages of samples are compared with the averages of the universes of which they are samples.

The two following illustrations (digits and latitudes) show how the distribution of a number of averages agrees with the normal curve of error.

In cases in which the theory applies, not only the standard deviation of the average can be given, but also the chances that the error of the average will exceed any given multiple of that standard deviation.

Since the universe is unknown it cannot always be stated whether its frequency group satisfies Edgeworth's conditions (p. 299) or not. We can sometimes test this from the samples themselves. Suppose that we take k samples each of n' items, and form their averages $\bar{x}_1, \bar{x}_2 \ldots \bar{x}_k$ into a frequency group. Then if the conditions in the universe are satisfied, this frequency group should be approximately normal, but not completely normal if n' is not large. If this is the case a major

sample may now be formed consisting of $n = n' \times k$ items. Its average equals the average of $\bar{x}_1,\ \bar{x}_2\ \ldots\ \bar{x}_k$, and as it ·is formed by selection of k things from a group, which, being approximately normal, satisfies the conditions in question, we may expect that the error in the major average has normal frequency with standard deviation $\dfrac{\sigma}{\sqrt{n}}$, where σ is calculated from the n observations. The k quantities $\bar{x}_1,\ \bar{x}_2\ \ldots\ \bar{x}_k$ should have for their standard deviation $\dfrac{\sigma}{\sqrt{n'}}$, approximately.

Thus in the example on p. 315 below the distribution of the 2000 items which are aggregated in 80 groups is not known. Here $k = 80$, $n' = 25$. The averages of the 80 groups are found to have standard deviation 1·628. It may be deduced that the standard deviation in the universe is approximately $1\cdot628 \times \sqrt{25} = 8\cdot14$. Then the standard deviation for the average based on the whole 2000 is $\dfrac{1\cdot628 \times \sqrt{25}}{\sqrt{2000}} = \dfrac{1\cdot628}{\sqrt{80}}$ as stated below.

As an alternative we can examine the frequency group formed by the n selections and see if it satisfies Edgeworth's conditions. If it does, we may take it that the error of the average has normal frequency.

Precision of Averages.

A sample was taken (as described on p. 281) of the house-holders' Census schedules in a number of districts, and the average number of persons per tenement was calculated in 12 districts.

Registration district.	Tene-ments.	Persons.	Persons per tenement.	Tenements.	Persons per tenement.	Standard deviation.
	In sample of 1 in 50.			In whole district.		
Bethnal Green, N.E. .	277	1,224	4·42	13,850	4·35	·14
S.W. .	278	1,261	4·54	13,905	4·60	·14
Shoreditch, S. .	187	792	4·24	9,331	4·26	·18
N.W..	152	693	4·56	7,623	4·34	·19
N.E. .	156	653	4·19	7,847	4·39	·19
Spitalfields . .	130	637	4·90	6,476	4·79	·21
Whitechapel .	117	519	4·44	5,914	4·72	·22
St. George . .	187	924	4·93	9,374	4·88	·18
Shadwell . .	95	387	4·07	4,800	4·37	·25
Limehouse . .	133	611	4·59	6,655	4·54	·21
Mile End, S.W. .	267	1,211	4·54	13,366	4·71	·15
N.E. .	207	839	4·05	10,364	4·40	·17

From a study of the Census volumes relating to the whole districts the standard deviations (σ) of the number of persons per tenement is found to range from 2·38 to 2·75.

The standard deviation for the first entry above is then $\dfrac{\sigma}{\sqrt{277}} = ·14$ if we take the lower and more stringent value of σ. The other standard deviations are calculated similarly.

The differences between the sample averages and the whole in 6 cases are less than the calculated standard deviation, in 4 cases exceed it by less than a quarter of itself, in 1 case by 30 per cent., and in 1 case the difference is twice the standard deviation.

Normal Distribution of Averages.

Ten digits were selected from successive final digits in seven-figure mathematical tables and summed, and the process repeated till 1000 totals were obtained.*

The average and the standard deviation of the group so obtained. were 45·014 and 9·205 respectively, as compared with 45 and $\sqrt{82·5} = 9·083$, which would be obtained from an indefinitely large random selection if the digits 0 to 9 were equally distributed.

The following table compares the distribution of the 1000 with the normal curve of error.

NUMBER OF TOTALS OF 10 FALLING WITHIN CERTAIN LIMITS.

Distance from average.	Calculated.	Standard deviation.	Observations.	Differences.
Above $\frac{1}{2}\sigma$ . . .	6	2·4	8	+ 2
2σ . . .	17	4·1	17	0
$\frac{1}{2}\sigma$ . . .	44	6·5	47	+ 3
·σ . . .	92	9·1	75	−17
$\frac{1}{2}\sigma$ . . .	150	11·3	157	+ 7
0 to $\frac{1}{2}\sigma$ . . .	191	12·4	197	+ 6
0 to $-\frac{1}{2}\sigma$ . . .	191	12·4	201	+10
$-\frac{1}{2}\sigma$ . . .	150	11·3	148	− 2
$-\sigma$ . . .	92	9·1	77	−15
$-\frac{1}{2}\sigma$ . . .	44	6·5	50	+ 6
-2σ . . .	17	4·1	20	+ 3
Below $-\frac{1}{2}\sigma$ . . .	6	2·4	3	− 3

The standard deviations are calculated from the formula $\sqrt{p(1-p)n}$ (formula (13)), where $n = 1000$ and p is the

---

* Such selections are found not to satisfy completely the conditions of independence. See *Statistical Journal*, 1912–13, p. 702 Nixon

proportion that falls within a grade in the normal law; thus between σ and $\frac{1}{2}\sigma$, ·092 of all are expected and $p = $ ·092.

In arranging the observations σ is taken as 9·205.

The differences between theory and observation are less than the standard deviation in 9 cases, and exceed it but are less than double in 3 cases.

The normal curve is therefore an adequate representation of the group.

The average as found from the whole sample of 10,000 is determined as 45·014 with standard deviation $\dfrac{9\cdot205}{\sqrt{1000}} = $ ·29, and is unexpectedly near the average of the sum of 10 digits in general.

From a geographical index containing 31,210 names, 25 were selected roughly at random, and their latitudes entered to a degree (ignoring minutes), the distinction between north and south being ignored. The 25 latitudes were then averaged and the process repeated till 80 averages were obtained.

The general average was 35·0° and the standard deviation of the group of 80 averages was 1·628°.

The following table compares the distribution with the normal curve of error on the same plan as in the last example.

Distance from average.			Calculated.	Standard deviation.	Observations.	Differences from nearest integer.
Above $\frac{1}{2}\sigma$	.	.	·5	—	1	0 or 1
2σ	.	.	1·3	?1·1	2	1
$\frac{3}{2}\sigma$	.	.	3·5	1·8	1	2 or 3
σ	.	.	7·4	2·6	10	3
$\frac{1}{2}\sigma$	.	.	12·0	3·2	12	0
0 to $\frac{1}{2}\sigma$	.	.	15·3	3·5	12	3
0 ,, $-\frac{1}{2}\sigma$	.	.	15·3	3·5	16	1
$-\frac{1}{2}\sigma$	.	.	12·0	3·2	15	3
$-\sigma$	.	.	7·4	2·6	7	0
$-\frac{3}{2}\sigma$	.	.	3·5	1·8	2	1 or 2
-2σ	.	.	1·3	?1·1	1	0
Below $-\frac{1}{2}\sigma$	.	.	·5	—	1	0 or 1

In eight cases the difference is below the standard deviation, and in the remaining two slightly above it.

The average 35·0°, as found from the sample of

$$80 \times 25 = 2000 \text{ latitudes,}$$

has standard deviation $\dfrac{1\cdot628}{\sqrt{80}}$ degrees $= $ ·18 degree, and is therefore not known accurately to the first decimal place.

Absolute Errors in Weighted Sums and Averages.

It has been shown above (p. 288) that if $H + E$ is the sum of n quantities selected independently from n frequency groups, whose averages are $_1\bar{u}, _2\bar{u} \ldots _n\bar{u}$, and standard deviations $\sigma_1, \sigma_2 \ldots \sigma_n$, and $H = {}_1\bar{u} + {}_2\bar{u} + \ldots + {}_n\bar{u}$, then $H + E_t$, the sum in any selection, $= {}_1u_t + {}_2u_t + \ldots + {}_nu_t$, has for its standard deviation s, where $s^2 = \sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2$

The same analysis can readily be re-arranged to show that, if we take a weighted sum

$H + E_t = W_1 \cdot {}_1u_t + W_2 \cdot {}_2u_t + \ldots + W_n \cdot {}_nu_t$, where $W_1, W_2 \ldots$ are constants, the standard deviation becomes

$$s^2 = W_1^2\sigma_1^2 + W_2^2\sigma_2^2 + \ldots + W_n^2\sigma_n^2 = S(W_t^2\sigma_t^2) \quad . \quad . \quad (55)$$

and the standard deviation, s_a, of the weighted average $\dfrac{H + E}{SW_t}$ is given by

$$s_a{}^2 = \frac{S(W_t^2\sigma_t^2)}{(SW_t)^2} \quad . \quad . \quad . \quad . \quad . \quad . \quad (56)$$

If n is large, so that $\dfrac{1}{\sqrt{n}}$ is negligible, and the other conditions stated on p. 299 are satisfied, the frequencies of the sum and average are normal, and the table on p. 271 can be used to ascertain the chances of deviations from the mean value H.

Let $\bar{\sigma}^2$ be the weighted mean value of $\sigma_1^2, \sigma_2^2 \ldots \sigma_n^2$, so that

$$\bar{\sigma}^2 S(W_t^2) = S(W_t^2\sigma_t^2)$$

Then $\qquad\qquad\qquad s^2 = \bar{\sigma}^2 S(W_t^2),$

and $\qquad\qquad\qquad s_a{}^2 = \bar{\sigma}^2 \dfrac{S(W_t^2)}{(SW_t)^2} \quad . \quad . \quad . \quad . \quad . \quad (57)$

Now let $SW_t = n\bar{w}$, and $W_t = \bar{w} + w_t$, and $n\sigma_w{}^2 = S(w_t^2)$, so that $\bar{w}$ and σ_w are the average and standard deviation of the W's regarded as a frequency group. $Sw_t = 0$.

Then

$$S(W_t^2) = S(\bar{w}^2 + 2\bar{w}w_t + w_t^2) = n\bar{w}^2 + 2\bar{w}Sw_t + Sw_t^2 = n(\bar{w}^2 + \sigma_w{}^2) \quad . \quad (58)$$

$$\text{and } s^2 = n\bar{\sigma}^2(\bar{w}^2 + \sigma_w{}^2) \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (59)$$

$$s_a{}^2 = \bar{\sigma}^2 \cdot \frac{n(\bar{w}^2 + \sigma_w{}^2)}{(n\bar{w})^2} \text{ and } \therefore s_a = \frac{\bar{\sigma}}{\sqrt{n}} \cdot \sqrt{\left(1 + \frac{\sigma_w{}^2}{\bar{w}^2}\right)} \quad . \quad . \quad (60)$$

The last formula gives in a convenient form the standard deviation of a weighted average, when the weights are known and not subject to error. The deviation of the original items is reduced in the ratio $1 : \sqrt{n}$, and becomes small when n is great, while the factor $\sqrt{\left(1 + \frac{\sigma_w{}^2}{\overline{w}^2}\right)}$ is rarely as great as $\sqrt{2}$, since $\frac{\sigma_w}{\overline{w}}$ measures the ratio of the standard deviation of the weights to their mean value, and this ratio in ordinary cases is less than unity.

If the average is unweighted, we have, of course,

$$s_a = \frac{\overline{\sigma}}{\sqrt{n}} \quad \ldots \ldots \ldots \ldots (61)$$

The fundamental formula $s^2 = S(W_t{}^2\sigma_t{}^2)$ was used by the Committee of the British Association on Small Incomes. (See *Statistical Journal*, 1910, December, p. 62, where different letters are employed.)

There were 31 classes in each of which the number not paying income tax was estimated as, say, N_t, with standard deviation s_t; their average income was I_t with standard deviation s'_t. The aggregate income of the class is then $N_t I_t$ with standard deviation σ_t, where

$$\sigma_t{}^2 = \text{Mean}\{(N_t + e_t)(I_t + e'_t) - N_t I_t\}^2$$
$$= \text{Mean}\{N_t e'_t + I_t e_t\}^2 = N_t{}^2 s'_t{}^2 + I_t{}^2 s_t{}^2,$$

when products of e's are neglected.

The standard deviation for the sum of $N_t I_t$ is therefore s, where

$$s^2 = S(N_t{}^2 s'_t{}^2 + I_t{}^2 s_t{}^2).$$

$s_1, s_2 \ldots$ and $s'_1, s'_2 \ldots$ were estimated separately for each class.

If we suppose that the numbers in the classes were known exactly and only the average incomes in the classes subject to error, then we should use the formula above $s^2 = S(W_t{}^2\sigma_t{}^2) =$ in this case $S(N_t{}^2 s'_t{}^2)$, which we of course also obtain by writing $s_t = 0$. The standard deviation of error in the average income of all the classes taken together is then $\dfrac{\sqrt{\{S(N_t{}^2 s'_t{}^2)\}}}{SN_t}$.

In the investigation $S(I_t{}^2 s_t{}^2) = 315 \times 10^5$, $S(N_t{}^2 s'_t{}^2) = 4 \times 10^8$, so that the errors in N were not important. $S(N_t) = 4023$, $S(N_t I_t) = 284,700$, and the average income in 1910 of persons

other than wage-earners not paying income tax may be written as £71 with standard deviation £5.

Relative Errors.

So far we have dealt with absolute errors and deviations, the actual differences between particular values and observations from their means or true values. It is now proposed to discuss relative errors and deviations (as used in Part I., Chapter VIII, *supra*).

If x is the observed value of a quantity whose true value or mean (as the case may be) is x', and $x = x'$ $(1 + e)$, then $e = \dfrac{x - x'}{x'}$ is the relative error or deviation.[*]

1. *Products and Quotients.*

If two factors F_1, F_2 are independent, and erroneously measured as $F_1(1 + e_1)$. $F_2(1 + e_2)$, and e is the resulting relative error in their product P, we have

$$P(1 + e) = F_1(1 + e_1) \cdot F_2(1 + e_2), \text{ where } P = F_1 F_2 \quad . \quad . \quad . \quad (62)$$
$$e = e_1 + e_2 + e_1 e_2 = e_1 + e_2, \text{ if products of } e\text{'s are negligible.}$$

Hence if σ, σ_1, σ_2 are the standard deviations of P, F_1, F_2, we have by the formula (34), p. 288, $\sigma^2 = \sigma_1^2 + \sigma_2^2$.

The result can be extended to any finite number of factors, so that

$$\sigma^2 = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \ldots \quad \quad . \quad . \quad . \quad (63)$$

The error of x^n, n finite, if given by $x^n(1 + e) = \{x(1 + e_1)\}^n$, where e_1 is the error of x.

$$\therefore \quad\quad\quad e = n e_1 + \frac{n(n-1)}{2} e_1^2 + \ldots = n e_1 \quad . \quad . \quad . \quad (64)$$

when squares are neglected, and the standard deviation of x^n is $n\sigma$ where σ is that of x.

The result is true when n is fractional. *E.g.* the error in a cube root is one-third the error of the quantity. Thus if a number 1006 is taken as 1000 (relative error ·006), the relative error in the cube root is ·002, the root being given as 10, instead of 10·02 $= 10$ $(1 + ·002)$ approx.

---

[*] In Chapter VIII above it was more convenient to take $\dfrac{x' - x}{x}$ as the error. If we call this e_1, the relation between the two is $e_1 = -e + e^2 - e^3 \ldots$, and $e_1 = -e$, when, as may generally be presumed, e^2 is negligible.

If e is the error in $Q = F_1/F_2$, and F_1 and F_2 are independent of each other,

$$Q(1+e) = \frac{F_1(1+e_1)}{F_2(1+e_2)}.$$

$$e = (1+e_1)(1+e_2)^{-1} - 1 = e_1 - e_2 + \text{squares and products.} \quad (65)$$

$$\sigma_q{}^2 = \sigma_1{}^2 + \sigma_2{}^2, \text{ where } \sigma_q \text{ is the standard deviation of } e. \quad (66)$$

If e is the error in a power, a^x, where a is known, and e_1 is the error in x

$$a^x(1+e) = a^{x(1+e_1)}$$

$$e = a^{xe_1} - 1 = e_1 \cdot x \log a \text{ when } e_1{}^2 \text{ is neglected} \quad . \quad . \quad (67)$$

Generally if e is the error in a function, $f(x)$

$$f(x) \times (1+e) = f\{x(1+e_1)\} = f(x) + e_1 x f'(x) + \ldots,$$

and $$e = x\frac{f'(x)}{f(x)} \cdot e_1 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (68)$$

2. *In Averages.*

Let $\overline{m}$ be the unweighted average of n quantities M_1, $M_2 \ldots M_t \ldots M_n$, and let $M_t = \overline{m} + m_t$, so that $Sm_t = 0$. Let $n\sigma_m{}^2 = Sm_t{}^2$.

Suppose the quantities erroneously observed as $M_t(1 + e_t)$ for M_t, etc., and let e be the relative error in their average.

$$\overline{m}(1+e) = \frac{1}{n} S(M_t(1+e_t)) = \overline{m} + \frac{1}{n} S(M_t e_t).$$

Then $$e = \frac{1}{n} S\left(\frac{M_t}{\overline{m}} e_t\right) \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (69)$$

If s_a, σ_t are the standard deviations of e, e_t then by formula (55) p. 316,

$$s_a{}^2 = S\left(\frac{M_t}{n\overline{m}} \sigma_t\right)^2 = \sigma^2 S\left(\frac{M_t}{n\overline{m}}\right)^2,$$

if σ^2 is the weighted mean of $\sigma_1{}^2 \ldots \sigma_t{}^2 \ldots$, or if all these standard deviations are equal

$$s_a{}^2 = \sigma^2 \cdot \frac{S(\overline{m} + m_t)^2}{n^2\overline{m}^2} = \sigma^2 \cdot \frac{\overline{m}^2 + \sigma_m{}^2}{n\overline{m}^2},$$

since $$Sm_t = 0,$$

and $$s_a = \frac{\sigma}{\sqrt{n}}\sqrt{\left(1 + \frac{\sigma_m{}^2}{\overline{m}^2}\right)} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (70)$$

This formula is of the same form as that relating to absolute errors in a weighted average. (Formula (60).)

In·this formula σ_m, $\bar{m}$ and $\sqrt{n}$ are known, and therefore the ratio of $\frac{s_a}{\sigma}$ can be stated exactly. σ_i has to be estimated from whatever circumstances are known about the individual measurements.

The conditions of p. 299 are generally satisfied when an average is computed, if the conditions of random sampling are preserved, and therefore the normal table of frequency is applicable if n is large ; it is approximately applicable if n is no greater than 20.

3. *In Weighted Averages.*

[Based on article in *Stat. Journal*, 1911-12, pp. 81-88.]

Let $\bar{m}_w = \dfrac{S(W_t M_t)}{SW_t}$, where M_t (and $\bar{m}$, σ_m) have the same meaning as before, and W_1, $W_2 \ldots W_t \ldots W_n$ are weights.

Let $W_t = \bar{w} + w_t$, where $n\bar{w} = SW_t$ and $Sw_t = 0$, and let $n\sigma_w^2 = Sw_t^2$.

Then $S(W_t M_t) = n\bar{w}\bar{m}_w$.

Suppose that the weights are imperfectly known, so that $W_t(1 + \eta_t)$ is taken instead of W_t etc.

Let the errors in the M's be as before, and let e be the resulting error in $\bar{m}_w$.

Then $\qquad \bar{m}_w(1 + e) = \dfrac{S\{W_t(1 + \eta_t) M_t(1 + e_t)\}}{S\{W_t(1 + \eta_t)\}}$.

$$\therefore e = \frac{S\{W_t(1 + \eta_t) M_t(1 + e_t)\} \cdot SW_t - S(W_t M_t) \cdot S\{W_t(1 + \eta_t)\}}{S(W_t M_t) \cdot S\{W_t(1 + \eta_t)\}}$$

$$= \frac{S(W_t M_t e_t) \cdot SW_t + S(W_t M_t \eta_t) SW_t - S(W_t \eta_t) \cdot S(W_t M_t)}{S(W_t M_t) \cdot SW_t},$$

$$\text{neglecting } e\eta \text{ and } \eta^2$$

$$= \frac{S(W_t M_t e_t)}{S(W_t M_t)} + \frac{S\{(W_t M_t \cdot n\bar{w} - W_t \cdot n\bar{w}\bar{m}_w) \eta_t\}}{S(W_t M_t) \cdot n\bar{w}}$$

$$= \frac{S(W_t M_t e_t)}{n\bar{w}\bar{m}_w} + \frac{S\{W_t(m_t + \bar{m} - \bar{m}_w) \eta_t\}}{n\bar{w}\bar{m}_w} \quad \ldots \ldots \quad (71)$$

Now $\bar{m}_w = \dfrac{S\{(\bar{w} + w_t)(\bar{m} + m_t)\}}{n\bar{w}} = \dfrac{n\bar{w}\bar{m} + \bar{m}Sw_t + \bar{w}Sm_t + Sw_t m_t}{n\bar{w}}$

$$= \bar{m}\left\{1 + \frac{1}{n}S\left(\frac{w_t}{\bar{w}} \cdot \frac{m_t}{\bar{m}}\right)\right\}$$

and $\bar{m}_w - \bar{m} = \dfrac{1}{n\bar{w}}S(w_t m_t)$ $\ldots \ldots \ldots \ldots \ldots$ (72)

$$e = \frac{S(W_t M_t e_t)}{n\bar{w}\bar{m}_w} + \frac{S(W_t m_t \eta_t)}{n\bar{w}\bar{m}_w}$$ approximately, if the difference $\bar{m} - \bar{m}_w$ is neglected, while in full the numerator of the second term is $S\{W_t (M_t - \bar{m}_w) \eta_t\}$.

Let $\bar{\sigma}$, σ_t, σ'_t be the standard deviations of e, e_t, η_t.

Then $\bar{\sigma}^2 = \dfrac{1}{(n\bar{w}\bar{m}_w)^2}\{S(W_t M_t \sigma_t)^2 + S[W_t (M_t - \bar{m}_w) \sigma'_t]^2\}$ . . (73)

Let $\sigma_1 = \sigma_2 = \ldots = \sigma$, and $\sigma'_1 = \sigma'_2 = \ldots = \sigma'$, or let σ^2, σ'^2 be weighted averages so that $\sigma^2 S(W_t M_t)^2 = S(W_t M_t \sigma_t)^2$ and $\sigma'^2 S(W_t m_t)^2 = S\{W_t (M_t - \bar{m}_w) \sigma'_t\}^2$.

Then $\bar{\sigma}^2 = \dfrac{1}{(n\bar{w}\bar{m}_w)^2}\{\sigma^2 . S(W_t M_t)^2 + \sigma'^2 S\{W_t (M_t - \bar{m}_w)\}^2\}$. (74)

σ and σ' must be estimated from whatever errors seem probable or possible in the circumstances of the measurement.

The other quantities involved can be calculated from the observations. A good approximation in ordinary cases to this result is

$$\bar{\sigma}^2 = \frac{\sigma^2}{n}\left(1 + \frac{\sigma_w^2}{\bar{w}^2}\right)\left(1 + \frac{\sigma_m^2}{\bar{m}^2}\right) + \frac{\sigma'^2}{n}\left(1 + \frac{\sigma_w^2}{\bar{w}^2}\right)\frac{\sigma_m^2}{\bar{m}^2} \quad . \quad (75)$$

This approximation is obtained as follows :—

$$S(W_t M_t)^2 = S\{(\bar{w}^2 + 2\bar{w}w_t + w_t^2)(\bar{m}^2 + 2\bar{m}m_t + m_t^2)\}$$
$$= n\bar{w}^2\bar{m}^2 + n\bar{w}^2\sigma_m^2 + n\bar{m}^2\sigma_w^2 + n\sigma_w^2\sigma_m^2$$
$$+ Sw_t^2(m_t^2 - \sigma_m^2) + 4\bar{m}\bar{w}Sw_t m_t + 2\bar{w}Sw_t m_t^2 + 2\bar{m}Sm_t w_t^2$$

$$\therefore \frac{S(W_t M_t)^2}{n(\bar{w}\bar{m})^2} = \left(1 + \frac{\sigma_w^2}{\bar{w}^2}\right)\left(1 + \frac{\sigma_m^2}{\bar{m}^2}\right) + \frac{\sigma_w^2\sigma_m^2 R_{22}}{\bar{w}^2\bar{m}^2} + \frac{4\sigma_w\sigma_m r}{\bar{w}\bar{m}} + \frac{2\sigma_w\sigma_m^2 r_{12}}{\bar{w}\bar{m}^2} + \frac{2\sigma_w^2 \sigma_m r_{21}}{\bar{w}^2\bar{m}}$$

where

$$r = \frac{Swm}{n\sigma_w\sigma_m}, \quad r_{12} = \frac{Swm^2}{n\sigma_w\sigma_m^2}, \quad r_{21} = \frac{Sw^2m}{n\sigma_w^2\sigma_m}, \quad R_{22} = \frac{Sw^2m^2}{n\sigma_w^2\sigma_m^2} - 1 = \frac{Sw^2(m^2 - \sigma_m^2)}{n\sigma_w^2\sigma_m^2}$$

$$S\{W_t(M_t - \bar{m}_w)\}^2 = S\{W_t(m_t + \bar{m} - \bar{m}_w)\}^2$$
$$= SW_t^2 m_t^2 + 2(\bar{m} - \bar{m}_w)SW_t^2 m_t + (\bar{m} - \bar{m}_w)^2 SW_t^2$$
$$= \bar{w}^2 . n\sigma_m^2 + 2\bar{w}Sw_t m_t^2 + Sw_t^2 m_t^2 - \frac{2}{\bar{w}}Sw_t m_t(2\bar{w}Sw_t m_t + Sw_t^2 m_t)$$
$$+ \left(\frac{Sw_t m_t}{n\bar{w}}\right)^2 n(\bar{w}^2 + \bar{\sigma}_w^2)$$

$$\therefore \frac{S\{W_t(M_t - \bar{m}_w)\}^2}{n(\bar{w}\bar{m})^2} = \frac{\sigma_m^2}{\bar{m}^2} + \frac{2\sigma_w\sigma_m^2}{\bar{w}\bar{m}^2} . r_{12}$$
$$+ \frac{\sigma_w^2\sigma_m^2}{\bar{w}^2\bar{m}^2}\left(R_{22} + 1\right) - 4\frac{\sigma_w^2\sigma_m^2}{\bar{w}^2\bar{m}^2}r^2 - 2\frac{\sigma_w^2\sigma_m^2}{\bar{w}^3\bar{m}^2}r . r_{21} + \frac{\sigma_w^2\sigma_m^2}{\bar{w}^2\bar{m}^2}\left(1 + \frac{\sigma_w^2}{\bar{w}^2}\right)r^2$$

$$\therefore \bar{\sigma}^2 = \frac{1}{n}\left(\frac{\bar{m}}{\bar{m}_w}\right)^2(l_1^2\sigma^2 + l_2^2\sigma'^2),$$

where

$$l_1^2 = \left(1 + \frac{\sigma_w^2}{\bar{w}^2}\right)\left(1 + \frac{\sigma_m^2}{\bar{m}^2}\right) + 4\frac{\sigma_w}{\bar{w}} . \frac{\sigma_m}{\bar{m}}r + 2 . \frac{\sigma_w}{\bar{w}} . \frac{\sigma_m^2}{\bar{m}^2}r_{12} + 2\frac{\sigma_w^2}{\bar{w}^2} . \frac{\sigma_m}{\bar{m}}r_{21} + \frac{\sigma_w^2}{\bar{w}^2} . \frac{\sigma_m^2}{\bar{m}^2}R_{22}$$

Y*

and

$$l^2 = \frac{\sigma_m{}^2}{m^2}\left\{\left(1 + \frac{\sigma_w{}^2}{w^2}\right) + \left(\frac{\sigma_w{}^4}{w^4} - 3\frac{\sigma_w{}^2}{w^2}\right)r^2 + 2\frac{\sigma_w}{w}r_{12} - 2\frac{\sigma_w{}^2}{w^3}r \cdot r_{21} + \frac{\sigma_w{}^2}{w^2}R_{22}\right\}$$

Now r, r_{12}, r_{21}, R_{22} each contain in their formulæ factors (m_t, w_t or $m_t{}^2 - \sigma_m{}^2$) whose sum is zero, and therefore, unless large values of the other factor ($m_t{}^2$, $w_t{}^2$) are found specially with positive or specially with negative values, the sum of the products is small, and terms containing these tend to be small in comparison with the other terms. Also $\dfrac{\overline{m_w}}{m} = 1 + r\dfrac{\sigma_w\sigma_m}{wm}$.

If we neglect r, r_{12}, r_{21}, R_{22} we obtain the approximation given above.

Examples.

Some examples, worked in detail, will show the relative magnitude of the quantities involved.

1. The first is a calculation of wages, where the weights are taken with great roughness, the number of persons of both sexes and all ages being taken for weights to compute the average wage of men only. It is only in very imperfect investigations that so deliberate an error would be introduced.

The contribution due to the errors in observations of quantities, typified by σ, has in the approximate formula two factors, each of which is always greater than 1 and generally less than 2; these factors can be computed from the observations.

On the other hand the contribution due to errors in weights, typified by σ' in (75), contains the factor $\left(\dfrac{\sigma_m}{m}\right)^2$ both in the approximate and in the complete formula, i.e. the square of the ratio of the standard deviation of the quantities to their mean value. In the cases, which are quite common when weighted averages are in question, where this ratio is small, the effect of errors in weights is smaller, and sometimes very much smaller, than the effect of equal errors in quantities. Hence the statements (pp. 94 and 185) that under ordinary conditions more attention should be paid to accuracy in quantities than to accuracy in weights.

Finally, as regards weighted averages, the table of probability on p. 271 may be applied to measure the chance of deviations greater than $\bar{\sigma}$, $2\bar{\sigma}$, $3\bar{\sigma}$ . . . if n is great, and it gives approximate values when n is as small even as 20.

METAL TRADES, EXCLUDING ENGINEERING AND SHIPBUILDING, 1906.

Cd. 5814, p. xi for numbers, and p. xiii for wages.

Trade.	Number of persons employed. W	Average earnings of men. M	
	ooo's.	s.	d.
Pig iron .	14	34	4
Iron and steel	54	39	1
Tinplate .	11	42	0
Railway carriages .	46	30	9
Iron castings .	12	31	4
Electric apparatus	15	34	7
Wire .	8	35	7
Brass, etc. .	8	31	9
Gold, silver, etc. .	8	36	6
Jewellery .	3	38	0
Edge tools .	3	31	2
Smelting .	8	31	5
Cycles .	7	34	4
Tubes .	7	28	3
Nails, etc. .	5	31	0
Bedsteads .	2	36	3
Farriery .	2	27	9
Scientific instruments .	2	36	10
Needles, etc. .	2	31	9
Chains, etc. .	1	35	4
Locks, etc. .	1	28	0
Watches and clocks	1	32	7
Typefounding .	1	33	3
Miscellaneous .	45	32	5
Total .	266		

n, the number of trades, $= 24$. S.W $= 266$. $\bar{w} = \frac{1}{n}$ S.W $= 11\frac{1}{12}$.

$\bar{m}$, the arithmetical average of the 24 entries of earnings, $= 33s.\ 6\frac{1}{8}d. = 33\cdot511s.$

$\sigma_m = 3\cdot47$. $\sigma_w = 14\cdot74$. m and w are the deviations of individual entries from $\bar{m}$ and $\bar{w}$.

$$r = \frac{Swm}{n\sigma_w\sigma_m} = \cdot150, \quad r_{12} = \frac{Swm^2}{n\sigma_w\sigma_m^2} = \cdot095, \quad r_{21} = \frac{Sw^2m}{n\sigma_w^2\sigma_m} = \cdot280,$$

$$R_{22} = \frac{Sm^2w^2}{n\sigma_m^2\sigma_w^2} - 1 = \cdot264. \qquad \left(\frac{\sigma_m}{\bar{m}}\right)^2 = \cdot011, \qquad \left(\frac{\sigma_w}{\bar{w}}\right)^2 = 1\cdot77,$$

$$\frac{\sigma_m}{\bar{m}} = \cdot104, \qquad \frac{\sigma_w}{\bar{w}} = 1\cdot33.$$

$\bar{m}_w$, the average of the earnings with the numbers given in the table as weights, $= 34s.\ 2\frac{1}{4}d., = \bar{m}\left(1 + \frac{r\sigma_m\sigma_w}{\bar{m}\bar{w}}\right); \quad \therefore \left(\frac{\bar{m}}{\bar{m}_w}\right)^2 = \cdot959.$

Y* 2

Then working with the notation of ·p. 321,

$$\therefore\ l_1{}^2 = 2{\cdot}77 \times 1{\cdot}011 + 4 \times 1{\cdot}33 \times {\cdot}104 \times {\cdot}150 + 2 \times 1{\cdot}33 \times {\cdot}011 \times {\cdot}095$$
$$+ 2 \times 1{\cdot}77 \times {\cdot}103 \times {\cdot}280 + 1{\cdot}77 \times {\cdot}011 \times {\cdot}264$$
$$= 2{\cdot}80 + {\cdot}083 + {\cdot}003 + {\cdot}102 + {\cdot}005 = 2{\cdot}99.$$

$l_1{}^2 \times \left(\dfrac{\overline{\overline{m}}}{\overline{m}_w}\right)^2 = 2{\cdot}87.$ The approximate formula gives $2{\cdot}80$.

$$l_2{}^2 = {\cdot}011 \{ 2{\cdot}77 + (3{\cdot}13 - 3{\cdot}99) \times {\cdot}0225$$
$$+ 2 \times 1{\cdot}33 \times {\cdot}095 - 2 \times 2{\cdot}35 \times {\cdot}150 \times {\cdot}280 + 1{\cdot}77 \times {\cdot}264 \}$$
$$= {\cdot}011 \{ 2{\cdot}77 - {\cdot}020 + {\cdot}253 - {\cdot}198 + {\cdot}467 \} = {\cdot}011 (2{\cdot}77 + {\cdot}50) = {\cdot}036.$$

$l_2{}^2 \times \left(\dfrac{\overline{\overline{m}}}{m_w}\right)^2 = {\cdot}035.$ The approximate formula gives $\cdot031$.

$$\overline{\sigma}^2 = \tfrac{1}{24}(2{\cdot}87\sigma^2 + {\cdot}035\sigma'^2).$$

The averages of the men's earnings in the separate trades are perhaps subject to an error of 6d. in 33s., in which case $\sigma = \tfrac{1}{66}$, $\sigma^2 = {\cdot}00023$.

The errors in the weights may be considerable, for the weights were deliberately taken as the whole number of persons instead of the number of men.

The error so introduced is computed from p. 10 of the report at about ·23, so that $\sigma'^2 = {\cdot}053$.

With these figures $\overline{\sigma}^2 = {\cdot}000027 + {\cdot}000077 = {\cdot}000104.$ $\overline{\sigma} = {\cdot}01$.

Hence the average may be written

$$\overline{m}_w(1 \pm \overline{\sigma})\ \text{or } 34s.\ 2\tfrac{1}{2}d.\ \pm 4d.$$

Though in this extreme case the error in the individual weights is taken 15 times as great as the error in the quantities, the resulting error is only ·0088 as compared with ·0052 due to quantities.

2. Perhaps the most important use of weighted· averages is as index-numbers of prices.

It was shown above (p. 204) that the change of the base year was equivalent to a change of weights ; such a change will by the theory used in this chapter produce an unimportant effect on the result, if the necessary conditions are found to hold.

Sauerbeck's numbers of the prices of commodities were tabulated for 1900 and 1911 and re-written with 1900 as base. Thus in the first entry the price of English wheat was 49 in 1900, 58 in 1911, when the average of the years 1867–77 is taken as 100. This was written 100 in 1900 and 118 in 1911. The 45 numbers so obtained give an arithmetical average

107·82, while the averages of the numbers as given by Sauerbeck in 1900 and 1911 were 75·07 and 79·69, whose ratio is 100 : 106·16.

In taking the simple average, when all the numbers are 100, we in effect give equal weights to the ratios ; whereas in Sauerbeck's setting, if $p_1, p_2 \ldots$ are the separate index-numbers in 1900 and $p_1', p_2' \ldots$ in 1911, the general index-numbers are $I_1 = \dfrac{p_1 + p_2 + \cdots}{45}$, $I_2 = \dfrac{p_1' + p_2' + \cdots}{45}$ and $I = 100 \dfrac{I_2}{I_1}$ gives the movement from 1900 to 1911, $i.e.$ 106·16,

$$I = 100\,\frac{Sp'}{Sp} = 100\,\frac{Sp \cdot \dfrac{p'}{p}}{Sp};$$ that is the ratios of the separate changes are weighted with the separate index-numbers in 1900.

We will examine the accuracy of the average on Sauerbeck's system, that is taking p, now written w, as a weight, and $\dfrac{p'}{p}$, now written m, as a quantity.

The quantities involved are the following :—

$$\bar{w} = 75\text{·}07, \quad \sigma_w = 20\text{·}67, \quad \frac{\sigma_w}{\bar{w}} = \cdot 275, \quad \bar{m} = 107\text{·}82, \quad \sigma_m = 20\text{·}03,$$

$\dfrac{\sigma_m}{\bar{m}} = \cdot 186, \quad r = - \cdot 2944, \quad \bar{m}_w = 106\text{·}2, \quad r_{12} = \cdot 506, \quad r_{21} = - \cdot 347,$ $R_{22} = \cdot 936$.

If we neglect r, r_{12}, r_{21}, R_{22}, as in formula (75),

$$\bar{v}^2 = \frac{\sigma^2}{45}(1\text{·}076)\,(1\text{·}035) + \frac{\sigma'^2}{45}(1\text{·}076) \times (\cdot 186)^2 = \sigma^2 \times \cdot 025 + \sigma'^2 \times \cdot 00083.$$

If we include these quantities $\bar{v}^2 = \sigma^2 \times \cdot 024 + \sigma'^2 \times \cdot 0012$.

The difference between the two is almost solely due to r_{12}, $i.e.$ to mean wm^2; abnormal increases from 1900 to 1911, measured by m, are on the whole found with abnormal movements from the base 1867–77 measured by w; but even this influence has not much effect.

The error, σ, in m is almost solely due to using round numbers, and tends to be about $\frac{1}{\sqrt{12}}\sigma$, and hence

$$\sigma^2 \times \cdot 024 = (\cdot 0005)^2,$$
and is negligible.

The error in w could be computed if we had a definite system of assigning importance to the commodities. In default of this, suppose they ought to have had equal weights,

as in "the alternative computation above. Then $\frac{\sigma_w}{\overline{w}} = \cdot 275$ measures the dispersion of the actual weights from the supposed true weights, and $\sigma'^2 \times \cdot 0012 = (\cdot 275 \times \cdot 034)^2 = (\cdot 0093)^2$.'

Hence $\sigma^2 = (\cdot 0005)^2 + (\cdot 0093)^2 = (\cdot 0093)^2$ approx., and the index-number may be written   .

$$106\cdot 2 \; (1 \pm \cdot 0093) = 106\cdot 2 \pm 1,$$

and this shows the kind of margin we should have in mind when using index-numbers.

Actually the difference between the numbers calculated on the two hypotheses is $107\cdot 8 - 106\cdot 2 = 1\cdot 6$.   ·   ·

Comparison of Averages.

If the errors in the two investigations are quite independent and lead to averages in the form $A_1 (1 \pm \sigma_1)$ $A_2 (1 \pm \sigma_2)$, the standard deviation of A_1/A_2 is $\sqrt{\sigma_1^2 + \sigma_2^2}$, by formula (66).

But it often happens that errors in the same· sense (both positive or both negative) are made in corresponding items at both dates ; thus the wages of a class may be underestimated at both dates. In such cases the error is reduced by the comparison.

Thus to take the case of a simple quotient $Q = F_1 \div F_2$.

If e_1 and e_2 are the relative errors in F_1, F_2 and their standard deviations are σ_1, σ_2, then the error in Q has standard deviation $\sqrt{(\sigma_1^2 + \sigma_2^2)}$ by formula (66).

But, if $d = e_1 - e_2$, mean $d^2 = $ mean $e_1^2 + $ mean $e_2^2 - 2$ mean $e_1 e_2$. The last term only vanishes if all values of e_2 are equally likely to occur with any value of e_1, and not if e_1 and e_2 are likely to be of the same sign.

E.g. if $e_2 = \frac{1}{2} e_1$ always, $\sigma_2^2 = \frac{1}{4} \sigma_1^2$, mean $e_1 e_2 = \frac{1}{2}$ mean $e_1^2 = \frac{1}{2} \sigma_1^2$, and mean $d^2 = \sigma_1^2 + \frac{1}{4} \sigma_1^2 - \sigma_1^2$ and the standard deviation of the ratio is $\frac{1}{2} \sigma_1$.

The necessary analysis for the ratios of weighted and unweighted averages is given in the Appendix, Notes 7 and 8.

The approximate formulæ are as follows, the notation being as on the previous page.

If s_r is the standard deviation of the ratio of two unweighted averages,

$$s_r^2 = \frac{1}{n} \sigma_d^2 \left(1 + \frac{\sigma_m^2}{\overline{m}^2}\right) \quad . \quad . \quad , \quad . \quad . \quad . \quad . \quad (76)$$

where σ_d is the standard deviation of the *difference* between e_t and e_t' the errors in measurements of the corresponding quantities M_t, M_t' at the two dates.

While if s_r is the standard deviation of the ratio of two weighted averages, then approximately under certain conditions

$$s_r{}^2 = \frac{1}{n}\left(1 + \frac{\sigma_w{}^2}{\overline{w}^2}\right)\left\{\sigma_d{}^2 + \left(\frac{\sigma_u}{1 + a}\right)^2\left(\sigma^2 + \sigma'^2\right)\right\} \quad . \quad . \text{ (77)}$$

where σ_d is as before, σ and σ' are standard deviations of the errors in quantities and weights, $M_t' = (1 + a + u_t)\,M_t$, where $Su_t = 0$, so that $1 + a$ measures the mean rate of growth of the quantities, and σ_u is the standard deviation of u and measures the scattering of the rates of growth.

If then the errors in the quantities tend to be the same at both periods, the first term in the bracket { } in (77) is small, and if the quantities grow at nearly the same rate the second term is small.   In any case the standard deviation diminishes with $\dfrac{1}{\sqrt{n}}$.

Under conditions which are often fulfilled it follows that very great accuracy can be obtained in the ratio of weighted averages, though the original errors in the measurement of quantities and in the systems of weighting are considerable. It is important not to vary the methods of computation, so as to obtain similar errors and a small value of σ_d.

Example.

DATA FOR ESTIMATING THE CHANGE IN AVERAGE WEEKLY WAGES IN CERTAIN
INDUSTRIES IN THE UNITED KINGDOM.

			1880.		1900.		
			Numbers W.	Wages M.	Numbers W'.	Wages M'.	Ratio of increase of M $1+a+u$
Agriculture :			ooo's.	shillings.	ooo's.	shillings.	
England and Wales	.	.	135	15	120	16·2	1·08
Scotland .	.	.	24	18	20	21·2	1·18
Ireland .	.	.	98	9	86	10·4	1·16
Building .	.	.	84	27	123	31·0	1·15
Printing .	.	.	8	31	13	32·9	1·06
Shipbuilding	.	.	7	28·5	13	34·8	1·22
Engineering	.	.	72	25	106	30·5	1·22
Coal .	.	.	44	23	75	34·3	1·49
Puddling .	.	.	9	31	11	38·1	1·23
Cotton .	.	.	52	16	54	19·5	1·22
Wool and worsted	.	.	12	14	12	13·6	·97
Worsted .	.	.	12	14	12	14·4	1·03
Gas .	.	.	3	27	8	31·0	1·15
Furniture .	.	.	12	23	18	24·8	1·08
			572		671		

The numbers are of all engaged in the industries from the General Report of the Census of England and Wales, Table 35. The rates of increase are from Mr. G. H. Wood's paper in the *Statistical Journal*, 1909, p. 93. The average wages are computed from various sources ; the accuracy of the ratios is more important than the accuracy of M.

$$n = 14,\ \bar{m} = 21\cdot54,\ \bar{m}' = 25\cdot20,\ \bar{m}_w = 18\cdot69,\ \bar{m}'_w = 24\cdot09,$$

$$\frac{\sigma_m}{\bar{m}} = \cdot319,\ \frac{\sigma_{m}'}{\bar{m}'} = \cdot351,\ \frac{\sigma_w}{\bar{w}} = 1\cdot00,\ \frac{\sigma_w'}{\bar{w}'} = \cdot90,\ \bar{u} = \cdot160,\ \sigma_u = \cdot12,$$

$$r = -\ \cdot42,\ r_{21} = -\ \cdot44,\ r_{12} = -\ \cdot42,\ R_{22} = \cdot25.$$

Ratio of unweighted averages $= \dfrac{\bar{m}'}{\bar{m}} = 1\cdot170$; of weighted averages $= \dfrac{\bar{m}'_w}{\bar{m}_w} = 1\cdot288.$

Mr. Wood gives $\frac{100}{80} = 1\cdot163$ and $\frac{100}{82} = 1\cdot219$ for these, using different weights.

$$s_r^2 = \frac{1}{14}\Big(1 + 1\cdot00^2\Big)\Big(\sigma_d^2 + \Big(\frac{\cdot12}{1\cdot160}\Big)^2(\sigma^2 + \sigma'^2)\Big)$$
$$= \cdot143\sigma_d^2 + \cdot0015(\sigma^2 + \sigma'^2),$$

by the approximate formula (77), and by the full formula (148), App.,

$$s_r^2 = \cdot145\sigma_d^2 + \cdot022\sigma^2 + \cdot0035\sigma'^2 + \cdot016\sigma'_d{}^2,$$

where σ'_d measures the difference between the errors in the weights at the two dates.

The approximate formula fails to do justice to the error in quantities owing to the great change in the weights in the period whose effect is ignored.

To see the effect of these errors, suppose the error in the wages in 1880 (σ) is $\frac{1}{20}$ and in the weights (σ') is $\frac{1}{10}$, and that similarity of error makes $\sigma_d = \frac{1}{2}\sigma$ and $\sigma_d' = \frac{1}{2}\sigma'$.

Then $s_r^2 = \cdot000091 + \cdot000055 + \cdot000035 + \cdot000040 = \cdot00022.$

$$s_r = \cdot015.$$

The ratio of the averages may be written

$$\frac{\bar{m}'_w}{\bar{m}_w}(1 \pm s_r) = 1\cdot288 \pm \cdot020$$

i.e. the percentage increase instead of being 29 may be anywhere from 27 to 31.

Actually the elemental. errors may be larger than• those here supposed. These figures are given as an example of method and to show• the influence of the various terms ; but $n \doteq 14$ is too small for the theory to be closely applicable, and a serious study of general wage-changes would need a wider range of industries and more exact determination of the numbers and average wages.

Significance of Differences Between Averages.

A very important problem that frequently arises in practical statistics is to determine whether the difference found between two averages of similar classes or groups could be due to the error incident to observation (especially to the inclusion of too small a number in a random sample) or can safely be attributed to real differences of characteristics. *E.g.*, if the observed death-rates of two classes are 14·7 and 14·3 per 1000, are we justified in saying that the death-rate of the first class is the higher, or should we expect a difference of ·4 if we simply separated two parts of the population arbitrarily ?

If the observed difference is greater than is to be expected in chance selection, it is said to be *significant, i.e.* significant of a real difference between the phenomena.

The general method of analysis is as follows : Suppose two classes containing n_1 and n_2 things yield averages $\bar{x}_1$ and $\bar{x}_2$. Calculate the standard deviation of the frequency curve of the differences between the averages of n_1 things and n_2 things selected indiscriminately from the whole universe from which the classes were segregated, and let this be σ.

Compare $\bar{x}_1 \sim \bar{x}_2$ with σ. The chance that the ratio is greater than 3 is ·0027, since the sum of the integrals of

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz \text{ from 3 to } \infty \text{ and } -3 \text{ to } -\infty,$$

$$\textit{i.e. } 2(\tfrac{1}{2} - F(3)) = 2(·5 - ·49865) = ·0027 \text{ (p. 271)}.$$

Similarly the chances that the ratio is greater than 2 or 1 are ·0456 or ·3174 ; and it is just as likely as not that the ratio is as great as ·674. If, then, $\bar{x}_1 \sim \bar{x}_2$ is not greater than ·674σ, there is no evidence of a real difference, that is, a difference due to the nature of the classes and not attributable to chance deviation. As $\bar{x}_1 \sim \bar{x}_2$ increases beyond this, the

improbability of the result as a chance event increases, till when the ratio equals 2 the odds are about 21 to 1 ($\cdot$9544 to $\cdot$0456) against. At 2σ we may say that the event is improbable unless the difference is real. At 3σ the odds against are about 370 to 1, and this is generally regarded as so improbable that the difference $\bar{x}_1 \sim \bar{x}_2$ is spoken of as significant. At 4σ the odds against are about 15,000 to 1. We can, of course, never arrive at certainty by this method ; we have rather to connect the word significant with the scale of probability. In the following paragraphs rules are given for calculating σ ; in every case the frequency group of the errors is normal, since the conditions described in previous sections are satisfied, and in every case the connection between σ and the probability of chance occurrence is that described in this paragraph.

A.—Cases of the Proportion of Things with Particular Characteristics in a Universe.

1. Let N be the number of things in a universe, of which pN have a particular characteristic where p and N are known. $q = 1 - p$.

Let n be selected at random, and $p'n$ be found to have the characteristic.

Then σ for $p' \sim p$ is $\sqrt{pq\left(\dfrac{1}{n} - \dfrac{1}{N}\right)}$ or $\sqrt{\dfrac{pq}{n}}$, if $\dfrac{n}{N}$ is negligible.

Example.—If dice are thrown 1200 times and 6 turns up 180 times, N $= \infty$, $p = \dfrac{1}{6}$, $n = 1200$, $p' = \dfrac{3}{20}$. $\sigma = \sqrt{\left(\dfrac{1}{6} \cdot \dfrac{5}{6} \cdot \dfrac{1}{1200}\right)} = \cdot 0108$

$$\frac{p - p'}{\sigma} = \frac{\cdot 0167}{\cdot 0108} = 1\cdot 6,$$

and there is an indication but no proof that the dice are not uniform in respect of their 6 faces.

2. Let two samples (n_1, p_1) (n_2, p_2) be selected from the universe, and neglect $\dfrac{n_1}{N}, \dfrac{n_2}{N}$.

The standard deviations of $p_1 - p$ and $p_2 - p$ are

$$\sqrt{\frac{pq}{n_1}} \text{ and } \sqrt{\frac{pq}{n_2}}$$

Hence the standard deviation of $p_1 \sim p_2 = (p_1 - p) \sim (p_2 - p)$

is the square root of the sum of the squares of their separate standard deviations (formula (34)) and

$$= \sqrt{\left\{ pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \right\}} \quad \dots \dots \quad (78)$$

If p is not known, but can only be deduced from the samples, the best value to take seems to be that found by merging the samples, viz. : $p(n_1 + n_2) = p_1 n_1 + p_2 n_2$.

Example.—In 1000 houses selected in a town, in 200 (n_1) the head of the household is an artisan, in 800 (n_2) a labourer. Children of school age are present in 80 of the first group $(p_1 = \cdot 4)$ and 420 of the second $(p_2 = \cdot 525)$. (The numbers are hypothetical.)

$$p \times 1000 = 80 + 420, \therefore p = \tfrac{1}{2} = q$$

$$\sigma \text{ for } p_1 \sim p_2 = \sqrt{\frac{1}{4}\left(\frac{1}{200} + \frac{1}{800}\right)} = \cdot 04$$

$$\frac{p_2 - p_1}{\sigma} = \frac{\cdot 525 - \cdot 4}{\cdot 04} = 3 \text{ approx.}$$

The difference is significant.

3. The samples $(n_1 p_1)$, $(n_2 p_2)$ are selected from different unknown universes, n_1 and n_2 being large.

E.g., suppose that out of 1000 men selected from two countries, 300 and 250 respectively are found to have blue eyes.

Here $p_1 = \dfrac{3}{10}$ in the selection, with $\sigma_1 = \sqrt{\dfrac{3 \times 7}{10^6}} = \cdot 014$, and the value for the whole country (if the selection had nothing to do with race or climate within the country) is approximately p_1. Similarly in the other country it is approximately $p_2 = \tfrac{1}{4}$.

The standard deviation for $p_1 \sim p_2$ is that for the difference between two independent groups, viz. :

$$\sqrt{\sigma_1^2 + \sigma_2^2} = \sqrt{\left(\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}\right)} = \cdot 02 \quad \dots \quad (79)$$

$$\frac{p_1 - p_2}{\sigma} = \frac{\cdot 3 - \cdot 25}{\cdot 02} = 2 \cdot 5.$$

This method is generally used when the death-rates of two occupational classes (*e.g.*, miners and bricklayers) are compared.

If of n_1, n_2 under observation m_1 and m_2 die in a year, the rates (r_1, r_2) are $\dfrac{m_1}{n_1} \times 1000$, $\dfrac{m_2}{n_2} \times 1000$, and $p_1 = \dfrac{m_1}{n_1}$, $p_2 = \dfrac{m_2}{n_2}$, since in the absence of other evidence it is assumed that the risk is the same throughout each class.

The miners are then assumed to be a random sample of a universe of miners, and similarly with the bricklayers.

Then σ for $r_1 - r_2$

$$= 1000 \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}} = 1000 \sqrt{\left\{ \frac{m_1 (n_1 - m_1)}{n_1^3} + \frac{m_2 (n_2 - m_2)}{n_2^3} \right\}} \cdot$$

A simpler procedure, however, is to compare each class with the adult male population as a whole. Then to find if the miners' death-rate differs from that of occupations in general we should use Case 1.

In the preceding it has been assumed that the chance p was the same throughout the universe. It may happen, however, that the universe consists of different regions or strata in which the chances are different, and the question arises whether we should proceed at random in the selection of a sample out of the universe as a whole or whether we should partially arrange the choice so as to take the same proportion out of each region or stratum. Mr. Yule (*Theory of Statistics*, p. 281) gives a formula which may be established as follows.

Let a universe contain $n_1, n_2 \ldots n_t$ things in t strata, and let the numbers which have a certain characteristic be $p_1 n_1, p_2 n_2 \ldots p_t n_t$ in these strata.

$N = n_1 + n_2 + \ldots + n_t$, and let $p_1 n_1 + p_2 n_2 + \ldots = PN$.

Let $k n_1, k n_2 \ldots k n_t$ be examined in the t strata, *i.e.* $kN = n$ in all.

Write $p_1 = P + d_1$, $p_2 = P + d_2 \ldots$,

where $P = p_1 \dfrac{n_1}{N} + p_2 \dfrac{n_2}{N} + \ldots$, so that $S(nd) = 0$.

The standard deviation of p_1 in the sample is

$$\sqrt{\frac{p_1 q_1}{k n_1}}$$

and similarly for p_2 etc.

Hence if σ is the standard deviation for P in the sample, by formula (55).

$$\sigma^2 = \left(\frac{n_1}{N} \right)^2 \frac{p_1 q_1}{k n_1} + \left(\frac{n_2}{N} \right)^2 \frac{p_2 q_2}{k n_2} + \ldots = \frac{1}{kN^2} (n_1 p_1 q_1 + n_2 p_2 q_2 + \ldots)$$

$$\therefore \; kN^2\sigma^2 = n_1 p_1 (1 - p_1) + \ldots$$
$$= S(np) - S(np^2) = NP - S\{n(P + d)^2\}$$
$$= NP - NP^2 - 2P \cdot Snd - Snd^2 = NPQ - N\sigma_p^2$$

where $\quad \sigma_p^2 = \dfrac{1}{N}(n_1 d_1^2 + n_2 d_2^2 + \ldots)$

$$\therefore \quad \sigma^2 = \frac{PQ}{n} - \frac{\sigma_p^2}{n} \quad \ldots \ldots \ldots \ldots \ldots \quad (80)$$

Here σ is the standard deviation for the observed result, P is the actual proportion in the universe, and σ_p^2 is the weighted mean square of the deviations in the strata.

If we took the numbers at random through the universe the standard deviation of the error would be σ_0, where $\sigma_0^2 = \dfrac{PQ}{n}$.

Hence $\sigma^2 = \sigma_0^2 - \dfrac{\sigma_p^2}{n}$, and by choosing proportionally from the various strata the standard deviation of the error involved is diminished.

In the investigation of the economic conditions of 4 towns (*Livelihood and Poverty*), instead of numbering all the houses and selecting 1 in 20 at random, we marked one out of every 20 throughout each street. By this means we secured that no district was completely unrepresented, which may possibly happen in a random selection, and we also got the advantage indicated by the formula just given, since social conditions in a street have a certain similarity. Suppose that there were 16,000 houses in 10 equal wards, and that in these wards the proportions below some assigned standard were ·02, ·06, ·10 ... ·38. Then $N = 16000$; $n_1 = n_2 = \ldots = 1600$; $p_1 = ·02$, $p_2 = ·06 \ldots P = ·2$; $d_1 = -·18$, $d_2 = -·14 \ldots$;

$$\sigma_p^2 = \tfrac{1}{10}(·18^2 + ·14^2 + \ldots), \; \sigma_p = ·115.$$

Now suppose 80 houses were examined in each ward, $n = 800$, $k = \tfrac{1}{20}$, $\sigma^2 = \dfrac{·2 \times ·8}{800} - \dfrac{·0132}{800}$, $\sigma = ·0136$, and the result may be written ·20 ± ·0136, or 20 ± 1·36 per cent.

In a non-stratified selection we should have had $\sigma = ·0141$. The gain in precision is very slight, but the method of selection by strata is in accordance with common sense and should be used where it is applicable.

B.—CASE OF A UNIVERSE CONTAINING A NUMBER OF MEASURABLE OBJECTS.

1. Let there be N objects in the universe, the average of whose measurements is $\bar{x}$ and standard deviation s.

n are selected at random and their average is found to be $\bar{x}_1$.

Then the standard deviation, σ, of $\bar{x}_1 \sim \bar{x}$ is $s\sqrt{\dfrac{1}{n} - \dfrac{1}{N}}$, by formula (52).

Example.—The average number of persons per tenement in a town of 10,000 tenements is 4·5, with standard deviation 2.

In 1000 working-class tenements the average is 4·7.

Here $N = 10,000$, $n = 1000$, $\bar{x} = 4\cdot5$, $\bar{x}_1 = 4\cdot7$, $s = 2$.

$$\sigma = 2\sqrt{\frac{1}{1000} - \frac{1}{10000}} = \cdot06, \qquad \frac{\bar{x}_1 - \bar{x}}{\sigma} = \frac{\cdot2}{\cdot06} = 3\cdot3.$$

2. The universe is only known by a sample, n, $\bar{x}$, s. A sub-sample of n_1 gives $\bar{x}_1$, σ_1.

Let n_2, $\bar{x}_2$, σ_2 be the residue, which, if the first sample were random and not of a class with a special average, would also be an independent random sample from the unknown universe.

Then $\qquad n_1 + n_2 = n, \quad n_1\bar{x}_1 + n_2\bar{x}_2 = n\bar{x}.$

$$\therefore \qquad \frac{\bar{x}_1 - \bar{x}_2}{n} = \frac{\bar{x}_1 - \bar{x}}{n_2} = \frac{\bar{x} - \bar{x}_2}{n_1}.$$

Also moments about the origin give

$$n(s^2 + \bar{x}^2) = n_1(\sigma_1^2 + \bar{x}_1^2) + n_2(\sigma_2^2 + \bar{x}_2^2).$$

Standard deviation for $\bar{x}_1 \sim \bar{x}_2$ is $\sqrt{\left(\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}\right)}.$

But the ratio of $\bar{x} \sim \bar{x}_1$ to $\bar{x}_1 \sim \bar{x}_2$ is constant and equal to $\dfrac{n_2}{n}$.

$\therefore$ standard deviation for $\bar{x} \sim \bar{x}_1$ is σ where

$$\sigma^2 = \frac{n_2^2}{n^2}\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) = \frac{\sigma_1^2}{n_1} + \frac{s^2 - 2\sigma_1^2}{n} - \frac{n_1}{nn_2}(\bar{x} - \bar{x}_1)^2,$$

as can be shown by eliminating σ_2 and x_2.

Let n_1 be less than n_2, and n and n_2 great.

Then $(\bar{x} - \bar{x}_1)\sqrt{n_1}$ is of order $\sigma\sqrt{n_1}$, *i.e.* of σ_1. Hence the term $(\bar{x} - \bar{x}_1)^2$ is negligible.

Hence $\qquad \sigma = \sqrt{\left\{\dfrac{\sigma_1^2}{n_1} + \dfrac{s^2 - 2\sigma_1^2}{n}\right\}}$ . . . . . (81)

See *Biometrika*, Vol. V., p. 182.

This method is used in the Scotch Census (Cd. 7163, p. 288) for comparing the size of families of men in different occupations.

If $\frac{n_1}{n}$ is small $\sigma \doteq \frac{\sigma_1}{\sqrt{n_1}}$, as it should from Case I when $\frac{n_1}{N}$ is neglected, and the observed σ_1 is taken for the unknown s.

If $\sigma_1 = s$, as will be the case if the standard deviation is not affected by class, but only the average affected,

$$\sigma = \sigma_1 \sqrt{\left(\frac{1}{n_1} - \frac{1}{n}\right)}$$ as was to be expected.

Example from the Scotch Census.

n, total number of marriages, $= 133,960$.

$\bar{x}$, average number of children per marriage, $= 5 \cdot 82$, with $s = 3 \cdot 099$.

Among boiler-makers, $n_1 = 923$, $\bar{x}_1 = 6 \cdot 00$, $\sigma_1 = 3 \cdot 039$.

$$\sigma = \sqrt{\left\{ \frac{9 \cdot 24}{923} + \frac{9 \cdot 60 - 18 \cdot 46}{133960} \right\}} = \cdot 10.$$

$\frac{\bar{x}_1 - \bar{x}}{\sigma} \doteq \frac{\cdot 18}{\cdot 10} = 1 \cdot 8$, which is barely significant.

Example.—Among the flour prices tabulated on p. 311 for U.S.A., 142 came from North Atlantic States.

	Number.	Average.	Standard Deviation.
U.S.A. . . .	$n = 267$	$\bar{x} = 2 \cdot 625$	$s = \cdot 293$
North Atlantic .	$n_1 = 142$	$\bar{x}_1 = 2 \cdot 748$	$\sigma_1 = \cdot 244$

$$\sigma = \sqrt{\left(\frac{\cdot 0595}{142} + \frac{\cdot 0858 - \cdot 1190}{267} \right)} = \cdot 017.$$

$\frac{\bar{x}_1 - \bar{x}}{\sigma} = \frac{\cdot 123}{\cdot 018} = 7$ approx., and the price in the North Atlantic States was definitely higher than the average for the whole country.

3. Two samples $(n_1 \bar{x}_1 \sigma_1)$ and $(n_2 \bar{x}_2 \sigma_2)$ are taken out of two known universes $(N_1 \bar{x}' s_1)$ and $(N_2 \bar{x}'' s_2)$.

σ for $\bar{x}_1 \sim \bar{x}_2$ is then

$$\sqrt{\left\{ s_1^2 \left(\frac{1}{n_1} - \frac{1}{N_1} \right) + s_2^2 \left(\frac{1}{n_2} - \frac{1}{N_2} \right) \right\}} = \sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)} \text{ approx.,}$$

since we have the difference between two independent observations, each coming under Case I.

If the universes are only known from samples, and $\frac{n_1}{N_1}$, $\frac{n_2}{N_2}$ are small we must take

$$\sigma^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}. \quad \ldots \ldots \ldots \quad (32)$$

There are some other variants, which can be treated on the same principles.

Example.—In food expenditures, similar to those tabulated on p. 310, in the whole group we have $\bar{x} = 10\cdot3$ (shillings), $s = 3\cdot3$.

$\bar{x}_1$ for the families of 566 skilled workmen was $10\cdot9$, and $\bar{x}_2$ for the families of 266 unskilled was $9\cdot3$.

The standard deviations for these groups were not calculated, but were probably nearly the same as s.

σ for $\bar{x}_1 - \bar{x}_2$ is $3\cdot3 \sqrt{(\frac{1}{566} + \frac{1}{266})} = \cdot25$.

$\dfrac{\bar{x}_1 - \bar{x}_2}{\sigma} = \dfrac{1\cdot6}{\cdot25} = 6$ approx., and the difference is significant.

The stratification of a universe of measurable objects is also treated by Mr. Yule (*Theory*, p. 345).

Let a universe ($N\bar{x}s$) be composed of groups $(n_1\bar{x}_1s_1)$, $(n_2\bar{x}_2s_2) \ldots$, and let kn_1, $kn_2 \ldots$ be selected from the groups, and the averages be found to be $(\bar{x}_1 + \delta_1)$, $(\bar{x}_2 + \delta_2) \ldots$, and the average of the kN to be $\bar{x} + D$; $kN = n$.

Then $\quad N = n_1 + n_2 + \ldots$; $\quad N\bar{x} = n_1\bar{x}_1 + n_2\bar{x}_2 + \ldots$

Write $\quad \bar{x}_1 = \bar{x} + d_1, \ \bar{x}_2 = \bar{x} + d_2 \ldots$;

then $\quad \mathrm{S}nd = 0$

$\quad Ns^2 = n_1(s_1^2 + d_1^2) + n_2(s_2^2 + d_2^2) + \ldots$

The squares of the standard deviations for δ_1, $\delta_2 \ldots$ are

$$\frac{s_1^2}{kn_1}, \ \frac{s_2^2}{kn_2} \ldots$$

Write σ for the standard deviation of D.

$$kN(\bar{x} + D) = kn_1(\bar{x}_1 + \delta_1) + kn_2(\bar{x}_2 + \delta_2) + \ldots$$

$\therefore \qquad\qquad D = \frac{n_1}{N}\delta_1 + \frac{n_2}{N}\delta_2 + \ldots$

$\therefore \qquad\qquad \sigma^2 = \left(\frac{n_1}{N}\right)^2 \frac{s_1^2}{kn_1} + \left(\frac{n_2}{N}\right)^2 \frac{s_2^2}{kn_2} + \cdots$

by formula (55).

$$= \frac{1}{kN^2}(n_1s_1^2 + n_2s_2^2 + \ldots).$$

Write σ_0 for the standard deviation of the average if the n samples had been taken at random from the universe as a whole.

Then $\sigma_0{}^2 = \dfrac{s^2}{n} = \dfrac{1}{n\mathrm{N}} \{n_1(s_1{}^2 + d_1{}^2) + n_2(s_2{}^2 + d_2{}^2) + \ldots\}$

$\therefore \qquad \sigma^2 = \sigma_0{}^2 - \dfrac{1}{n} \cdot \dfrac{S(nd^2)}{\mathrm{N}}.$

Write $\mathrm{N}\sigma_m{}^2 = S(nd^2)$, so that $\sigma_m{}^2$ is the weighted mean square of the deviations of the averages in the strata.

Then $\qquad\qquad\qquad \sigma^2 = \sigma_0{}^2 - \dfrac{\sigma_m{}^2}{n} \quad \ldots \ldots \ldots \; (83)$

The precision of the average is improved by stratification, as in the previous case (formula (80)).

Thus in the example on p. 313, let N be the total number of tenements in the last seven districts named (Spitalfields and onwards). $\bar{x}$, the average number of persons per tenement in the districts combined, is found from the Census to be 4·64, with $s = 2\cdot75$. $k = \frac{1}{50}$, since one tenement in 50 was recorded in the selection; N = 57000, and $n = 1140$.

	Number of tenements.	Persons per tenement. $\bar{x}$
Spitalfields	$n_1 = 65^{00}$	$+\cdot15 = d_1$
Whitechapel	$n_2 = 59$	$+\cdot08 = d_2$
St. George	$n_3 = 94$	$+\cdot24 = d_3$
Shadwell	$n_4 = 48$	$-\cdot27 = d_4$
Limehouse	$n_5 = 66$	$-\cdot10 = d_5$
Mile End, S.W. . . .	$n_6 = 134$	$+\cdot07 = d_6$
,, N.E. . . .	$n_7 = 104$	$-\cdot24 = d_7$
	570	

$S(nd^2) = 1806 \qquad \sigma_m{}^2 = \dfrac{1806}{57000} = \cdot0317$

$\sigma_0{}^2 = \dfrac{(2\cdot75)^2}{1140} = \cdot006634 \qquad \sigma_0 = \cdot08145$

$\sigma^2 = \cdot006634 - \dfrac{\cdot0317}{1140} = \cdot006606 \qquad \sigma = \cdot08128$

The improvement obtained by sampling in the seven strata represented by the districts is very slight.

Existence of a Trend.

Further applications of the same principles are made when we consider a time-series of observations and examine whether the fluctuations and movements are random or show the existence of a trend or of periodicity.

z*

Th̀e method, and its difficulties, can be shown sufficiently by two examples.

1.—The Recorded Times for " The Oaks " from 1850 to 1899 are as Shown Below

	min. sec.		min. sec.		min. sec.		min. sec.		min. sec.
1850	2 56	1860	2 56	1870	2 52	1880	2 49	1890	2 40⅛
1851	2 52	1861	2 44	1871	2 51	1881	2 46	1891	2 54⅜
1852	3 0	1862	2 49	1872	2 52	1882	2 49	1892	2 43½
1853	2 52	1863	2 54	1873	2 50⅜	1883	2 53	1893	2 44⅜
1854	3 0	1864	2 47	1874	2 48½	1884	2 49	1894	2 50
1855	2 58	1865	2 51	1875	2 49½	1885	2 .43¾	1895	2 48¾
1856	3 4	1866	2 53	1876	2 50	1886	2 54¼	1896	2 45¼
1857	2 50	1867	2 54	1877	2 54½	1887	2 50⅜	1897	2 45
1858	2 53½	1868	2 47½	1878	2 54	1888	2 42½	1898	2 45¼
1859	2 55	1869	2 59	1879	3 2	1889	2 45	1899	2 44
Ten yearly average	2 56·05		2 51·45		2 52·395		2 48·22		2 46·26

These figures fit fairly well a normal curve with average 2 min. 50·87 secs. and standard deviation 5·20 secs. The standard deviation for the difference between two records is therefore $5·2\sqrt{2} = 7·4$ secs. This is only exceeded eleven times between consecutive years, and no difference between consecutive years reaches twice this ; hence there is no proof of any sudden change having taken place between two races. The difference between some of the times for years early in the period and those later in some cases exceeds 20 seconds. The standard deviation for the difference between the averages for two periods of ten years is $5·2\sqrt{\left(\dfrac{1}{10}+\dfrac{1}{10}\right)} = 2·33$ secs. The difference between the averages for 1850–9 and 1890–9 is nearly 10 seconds, and is significant, as is the difference between the averages for 1850–9 and 1880–9. The intermediate differences are hardly significant. Hence we find that some cause was at work which gradually quickened the race between the fifties and the eighties.

2.—The Marriage Rates for England and Wales from 1860 to 1909 were

1860	17·1	1870	16·1	1880	14·9	1890	15·5	1900	16·0
1861	16·3	1871	16·7	1881	15·1	1891	15·6	1901	15·9
1862	16·1	1872	17·4	1882	15·5	1892	15·4	1902	15·9
1863	16·8	1873	17·6	1883	15·5	1893	14·7	1903	15·7
1864	17·2	1874	17·0	1884	15·1	1894	15·0	1904	15·3
1865	17·5	1875	16·7	1885	14·5	1895	15·0	1905	15·3
1866	17·5	1876	16·5	1886	14·2	1896	15·7	1906	15·7
1867	16·5	1877	15·7	1887	14·4	1897	16·0	1907	15·9
1868	16·1	1878	15·2	1888	14·4	1898	16·2	1908	15·1
1869	15·9	1879	14·4	1889	15·0	1899	16·5	1909	14·7
Ten yearly average	16·70		16·33		14·86		15·56		15·55

The average for 50 years is 15·80, and the standard devia-tion for the 50 records is ·89, and, taken irrespective of order, the distribution is nearly normal.   There are no sudden jumps from one year to the next.   The standard deviation for the difference between two averages of ten years is ·4, and hence the fall from 1870–9 to 1880–9 and the subsequent rise is significant.

The first twenty-five years shows greater variation than the second twenty-five, and we can make a finer test.

1860–1884. $\sigma = $·894 $\sigma\sqrt{\tfrac{2}{t}} = $·566.		1835–1909. $\sigma = $·613 $\sigma\sqrt{\tfrac{2}{t}} = $·388.	
	Average.		Average.
1860–4	16·70	1885–9	14·50
1865–9	16·70	1890–4	15·24
1870–4	16·96	1895–9	15·88
1875–9	15·70	1900–4	15·76
1880–4	15·22	1905–9	15·34

There is a significant fall from 1870–4 to 1885–9 and a significant rise from 1885–9 to 1895–9.

The argument should be illustrated by a diagram, which will suggest to what periods the test should be applied.

Periodicity.

The general question of the existence of a period of a length not predetermined is a mathematical problem, that of harmonic analysis, and is not suitable for discussion here ; but we can test the influence of periodicity if the length of the period is given.

Take the case of a given interval, say one year where the records are monthly, and consider whether the differences between, say, January and February are such as might occur in a random choice of observations irrespective of time. Suppose the records extend over t years, so that there are $12 \times t$ in all, that their average is $\bar{x}$ and their standard deviation from the average $\sigma = \sqrt{\left(\dfrac{S(x - \bar{x})^2}{12t}\right)}$, where x stands for any obser-vation.

The standard deviation of the difference between two averages each of t records selected at random is

$$\sqrt{\left\{\frac{\sigma^2}{t} + \frac{\sigma^2}{t}\right\}} = \sigma\sqrt{\frac{2}{t}},$$

and the chance of exceeding this deviation is found from the table of normal probability (p. 271), if the records regarded as a group are nearly normal, or otherwise satisfy the conditions of p. 299. If instead of taking random selections we compare the average of the t January records with that of the t February records and find that the difference exceeds twice or three times $\sigma \sqrt{\dfrac{2}{t}}$, and similarly for other months, then we have evidence that the quantities measured are affected by the time of year, unless the records of a month include some quite abnormal entry.

It is not, however, easy in this method to include all the evidence. Thus if we take the 180 records of unemployment on p. 161 we find that the average is 4·269 and the standard deviation is 1·924. The standard deviation for the difference between two averages of 15 is therefore $1\cdot924 \sqrt{\dfrac{2}{15}} = \cdot70$. This is exceeded, but not greatly, when we compare the averages for January or December with those for April, May, June or July, and there is no other difference which might not arise in random selection. There is, however, cumulative evidence which can hardly be measured. Thus the averages fall from December through January, February, March (if we omit the abnormal entry in 1912), and April to May, and rise month by month from May to October. This suggests a wave motion which the method here suggested is incapable of measuring.

Another method, also difficult to make precise, is to compare the numbers of falls and rises from an assigned month to the next.

	Jan. to Feb.	Feb. to Mar.	Mar. to Apr.	Apr. to May.	May to June.	June to July.	July to Aug.	Aug. to Sept.	Sept. to Oct.	Oct. to Nov.	Nov. to Dec.	Dec. to Jan.
Falls	12	13	10	10	5½	7	2	7½	8½	10	15	10
Rises	3	2	5	5	9½	8	13	7½	6½	5	0	4

Thus in 12 years the February number was less than that for the preceding January, and in 3 years it was greater. Where the numbers are equal, ½ is counted for each row. Now in 15 trials in each of which + and − are equally likely, the chance of obtaining 10 or more of like sign is about ⅓, so that the movements March to April, April to May, October to November are not very improbable in a random selection. The chance of obtaining 12 or more of the same sign is only $\frac{1}{16}$,

and the movements from · January to February, February to March, July to August, November to December, would hardly occur, if there were no influence from the season.

The conclusion seems to be that there is a cumulative decrease from November to March or April or May, and a cumulative rise during the early summer.

Another example gives more definite results. The records of the catch of haddocks are recorded (*North Sea Fisheries Investigation*, Granton) monthly for 18 years, the unit being 1 cwt. per month per vessel. The average is 172 and the standard deviation of the 216 records is about 108.

The standard deviation for the difference between the average of one month compared with the average of all is $108 \sqrt{\left(\frac{1}{18} + \frac{1}{216}\right)} = 26 \cdot 5$ approx., and for the difference between the averages of two months is $108 \sqrt{\frac{2}{18}} = 36$, in both cases if the selections were random and there were no seasonal influence.

The averages recorded are :—

January	. 101	April	. 83	July	. 247	October	. 227
February	. 115	May	. 145	August	. 282	November	. 181
March	. 125	June	. 196	September	. 267	December	. 101
						Year	. 172

Here January, February, April and December are more than twice 27 below the average, March and May are not less than 27 below the average, each month from July to October is more than twice 27 above the average, June and November are within 27 of the average. The conclusion is definitely that the season July to October is better than the season December to April.

Also the movements between consecutive months are more than 36 in the following cases : March to April, April to May, May to June, June to July, September to October, October to November, and November to December. April is clearly the worst month, but it is doubtful whether August is established as the best.

From the original figures we have

Number of	Jan. to Feb.	Feb. to Mar.	Mar. to Apr.	Apr. to May.	May to June.	June to July.	July to Aug.	Aug. to Sept.	Sept. to Oct.	Oct. to Nov.	Nov. to Dec.	Dec. to Jan.
Falls	5½	10	16	4	4	7	4	11	11	11	15	8
Rises	12½	8	2	14	14	11	14	7	7	7	3	9

Hére a number greater than 11 or less than 7 is likely to be significant.

NOTES.

1. The standard deviation of an average is often given as $\dfrac{\sigma}{\sqrt{n-1}}$, instead of $\dfrac{\sigma}{\sqrt{n}}$ as on p. 289, on the ground that we should distinguish between the deviation from the unknown true average and that from the average of observations.

Let $\bar{x}_0$ be the true average of a group whose standard deviation is σ_0, and let n things be selected from it which give an average $\bar{x}$, and which separately are X_1, $X_2 \ldots$ with standard deviation σ.

Write $\bar{x} = \bar{x}_0 + d$.

The deviations of X_1, $X_2 \ldots$ from the true average are $X_1 - \bar{x}_0$, $X_2 - \bar{x}_0 \ldots$, and the standard deviation of these is by hypothesis σ_0.

Hence $\sigma_0{}^2 =$ Mean $(X - \bar{x}_0)^2 =$ Mean $(X - \bar{x} + d)^2 =$ Mean $(X - \bar{x})^2 + d^2$

$$= \sigma^2 + \frac{\sigma_0{}^2}{n}, \text{ since from formula (38) the standard deviation of the}$$

average is $\sigma_0/\sqrt{n}$

$$\therefore \; \sigma_0{}^2 = \frac{n}{n-1}\sigma^2$$

and

$$\frac{\sigma_0}{\sqrt{n}} = \frac{\sigma}{\sqrt{n-1}} = \sqrt{\left\{\frac{S(X-\bar{x})^2}{n(n-1)}\right\}}$$

Hence the observed σ should be divided by $\sqrt{n-1}$, not $\sqrt{n}$.

The modification is only of theoretic importance, for the difference is only perceptible with quite small values of n, and σ is liable to an error of the same order as this difference in any case.

2. When as on p. 159 and pp. 375, 387 we measure the deviation of an observation in a time series from the average of t years of which it is the centre, we ought to pay attention to the risk of error due to averaging, measured by $\sigma/\sqrt{t}$, where σ is the standard deviation of the observations in neighbouring years. The standard deviation of the difference between an observation and such an average is not σ but $\sqrt{\left(\sigma^2 + \dfrac{\sigma^2}{t}\right)} = \sigma\sqrt{\dfrac{t+1}{t}}$. Since t is small, the error is perceptible, and the deviations as shown on such a diagram as that facing p. 155 are imperfectly estimated, and the measurement of correlation on pp. 386–7 lacks precision.

CHAPTER V.

EMPIRICAL FREQUENCY EQUATIONS.

It cannot be assumed that frequency groups in general are expressible by the law of great numbers, for the particular complex of independent causes which leads to its equation cannot be postulated for observational groups in general. The main use of the normal curve is in its application to averages or other functions whose methods of generation are known. Its applicability to anthropometrical or biometrical groups must be verified for each class of measurements, and the question whether mental and moral characteristics are normally distributed needs special investigation. There is, however, a presumption that in very many classes the normal distribution represents fairly the central portion of a group (from the centre to once or twice the standard deviation) and that the chance of an observation differing from the average by more than twice the standard deviation is not large, and consequently the table of normal frequency affords some guidance even in non-normal cases.

For complete description of groups either a more elastic system is needed to include wider classes than are covered by the curve of error, or equations on an empirical basis should be found to fit special classes of observations. In this chapter we deal very briefly with equations that serve one or other of these purposes.

The general method is to select a mathematical equation involving 2, 3 or 4 unknown constants, the constants being so chosen as to make the curve represented by the equation fit the diagram formed from the observations ; the number of points on the diagram being more numerous than the number of constants, ,we obtain more equations than unknowns and the best solution has to be chosen. A usual way of meeting

such a difficulty is by the method of least squares (p. 452), but with observational frequency curves Professor Pearson's method is generally used, equating the moments deduced mathematically from the equation of the curve to the moments obtained (as in Chap. I, p. 253) from the observations. This method has already been used (p. 305) when the average, standard deviation, and skewness ($\bar{x}$, σ, κ) have been obtained from the first three moments of the observations, and it is always used in the system described in the next paragraph. Other methods are to obtain those constants which satisfy the condition that the observations would be found in a random sample with minimum improbability or to select a small number of chosen points at which the equations shall be exactly satisfied.

Professor Karl Pearson's System.

It is necessary to call attention to the system of curves introduced by Professor Karl Pearson, since the notation involved has become general in statistical investigations, and it is advisable to indicate their relationship to the present treatment. For a detailed treatment, however, the reader is referred to Mr. Elderton's book, *Frequency Curves and Correlation*, and Mr. Hardy's *Theory of the Construction of Tables of Mortality*.

μ_0, μ_1, ... μ_t ... are used to denote the successive moments of a frequency curve. μ_0, the area, is taken as unity. μ_1 is zero, if the curve is referred to the ordinate through the centre of gravity of the curve. When this is the case, σ, the standard deviation, is defined as $\sqrt{\mu_2}$. β_1 is written for $\dfrac{\mu_3^2}{\mu_2^3}$ and β_2 for $\dfrac{\mu_4}{\mu_2^2}$.

The equation

$$D_x y = \frac{(x + a)y}{b_0 + b_1 x + b_2 x^2} \quad . \quad . \quad . \quad . \quad (84)$$

is the basis of the analysis.

This satisfies the condition that the curve should touch the axis when $y = 0$ and also be horizontal at one other position, namely when $x = -a$. That is, the curve has one mode.

It is found in practice that it is useless to continue the denominator after the term b_2x^2.

The integration of this equation leads to the three alternative general forms:

$$y = y_0\left(1 + \frac{x}{a_1}\right)^{\nu a_1}\left(1 - \frac{x}{a_2}\right)^{\nu a_2}, \; y = y_0\left(1 + \frac{x^2}{a^2}\right)^{-m} e^{-\nu \tan^{-1}\frac{x}{a}},$$

and

$$y = y_0(x - a)^{q_2} x^{-q_1},$$

where y_0 and the sets of three constants (ν, a_1, a_2), (m, a, ν), (a, q_2, q_1), are determinable by means of moments from a, b_0, b_1, b_2 in the basic equation. Mr. Hardy gives an alternative method of analysis based on an apparently simpler notation.

When there are special values of a, b_0, b_1, b_2 or special relations between them, simpler equations involving only two constants, or even only one, are obtained. In all, seven principal types are distinguished, and Mr. Elderton shows how each can be fitted to appropriate observational frequency groups. The algebra and the arithmetic involved are somewhat heavy. The results of the application of the method to food expenditure are given on p. 310.

The equation of the normal curve of error can be written in the form

$$D_x y = -\frac{xy}{\sigma^2} \quad \ldots \ldots \ldots (85)$$

and is one of the special types.

The second approximation to the general curve of error gives

$$D_x y = -\frac{\left(x + \frac{\kappa\sigma}{2}\right)y}{\sigma^2 + \frac{\kappa\sigma \iota}{2}} \quad \ldots \ldots \ldots (86)$$

where κ^2 is neglected, and is also a special case.

It has been found, especially by Professor K. Pearson and his co-workers, that unimodal observational frequency groups can very generally be represented adequately by one or other of the variants of the formula. Hence the calculation of the average, σ, β_1, and β_2 from the observations forms a general and useful way of expressing a group by four intelligible quantities, carrying further the process by which an average is commonly taken as representing a group.

From these quantities the equation of the curve representing the group can be deduced in its appropriate form, and then it is possible to interpolate values of y for any value of x, whatever the grading of the observations may be.

It is not proposed here to discuss how far these equations can be used in questions of probability, nor to consider how far the fundamental formula is empirical and how far it is dependent on hypotheses of chance generation.

Professor Edgeworth's Method.

Professor Edgeworth has developed a formula based on a transformation of the normal curve of error which represents classes of cases whose skewness is too great to allow them to be included under the second approximation of the generalised law of error. It has not yet been tried sufficiently to decide how far it is useful for description, interpolation, or other purposes. (See *Statistical Journal*, two series of papers, commencing December, 1898, and July, 1916, respectively.)

Professor Pareto's Equation.

The equation $D_x y = -\dfrac{my}{x}$, obtainable from the system described above by taking the case where $b_0 = 0$ and $b_2 = \dfrac{b_1}{a} = -\dfrac{1}{m}$, represents a curve which slopes downwards to the right for all possible values of x, when m is positive.

In its integral form it is $\log y = -m \log x + \text{const.}$, or $y = Cx^{-m}$.

The area of the curve from x to ∞ is

$$z = \int_x^\infty Cx^{-m}dx = \left[\frac{Cx^{1-m}}{1-m}\right]_x^\infty$$

$$= \frac{C}{(m-1)x^{m-1}}, \text{ if } m > 1.$$

Write a for $\overline{m-1}$ and A for $\dfrac{C}{m-1}$, and we have

$$y = \frac{Aa}{x^{a+1}}, \quad z = \frac{A}{x^a} \quad \ldots \ldots \quad (87)$$

The last equation is the simplest form of "Pareto's Law" for incomes. Here A and a are constants and z is the aggregate

number of persons whose incomes are at or above $£x$ (or x francs, etc.).

Other groups, e.g. the number of houses of various annual values, where the number of instances and the variable are capable of very wide ranges of values, and which are suitable for graphing on double logarithmic scales, are also found to conform to the same formula.

In the case of incomes, the aggregate of incomes from $£x_1$ to $£x_2$ is

$$\int_{x_1}^{x_2} xy\,dx = \frac{Aa}{a-1}\left(\frac{1}{x_1^{a-1}} - \frac{1}{x_2^{a-1}}\right)$$

The number of incomes in the range is $A\left(\dfrac{1}{x_1^a} - \dfrac{1}{x_2^a}\right)$.

The law is not generally found applicable to very low or very high incomes.   If it did extend to the maximum income, we should have

Aggregate income at or above $£x = \dfrac{Aa}{(a-1)x^{a-1}}$,

Number of incomes at or above $£x = \dfrac{A}{x^a}, = N$, say,

and hence

Average income from $£x$ upwards $= \dfrac{a}{a-1}\,.\,x$,

and these equations would give A and a immediately from records of incomes.

Pareto's equation fits the statistics of incomes of 1911–12 paying super-tax very well over the range £5,000–£55,000 ; above the latter income it gives numbers in excess of the recorded income.

$a = 1.5$, $\log A = 9.618$ are found to give a close fit.

Range of Incomes (ooo's).				Number of Incomes.	
				Calculated.	Recorded.
£5 to £10	.	.	.	7,546	7,411
10 ,, 15	.	.	.	1,890	2,029
15 ,, 20	.	.	.	790	787
20 ,, 25	.	.	.	424	438
25 ,, 35	.	.	.	411	382
35 ,, 45	.	.	.	199	186
45 ,, 55	.	.	.	103	107
55 ,, 65	.	.	.	70	56
65 ,, 75	.	.	.	50	37
75 ,, 100	.	.	.	118	55
100 and over	.	.	.	83	66
Totals .	.	.	.	11,700	11,554

Aggregate of incomes : Calculated, £166,000,000 ; recorded, £145,000,000.

If a doubly-logarithmic diagram is drawn, the range over which a straight line is a good approximation can be seen, and a trial value of a is suggested by its gradient. This value may be tested by choosing two values for x, say x_1 and x_2, which give values of N represented by points lying nearly on the empirical line, say N_1 and N_2.

Then
$$a = \frac{\log N_1 - \log N_2}{\log x_2 - \log x_1}$$

If we take $x_1 = 5{,}000$, $x_2 = 45{,}000$ from the table just given, we have $N_1 = 11{,}554$ and $N_2 = 321$; hence $a = 1{\cdot}63$.

This method, however, assumes that the number up to the maximum income conforms to the law, which is not generally the case; and in practice it is better to take three values x_1, x_2, x_3.

Then the equation

$$f(a) = \left(\frac{1}{x_1{}^a} - \frac{1}{x_2{}^a}\right) \div \left(\frac{1}{x_1{}^a} - \frac{1}{x_3{}^a}\right).$$

$$= \frac{\text{number of incomes from } x_1 \text{ to } x_2}{\text{number of incomes from } x_1 \text{ to } x_3} = k, \text{ a known quantity,}$$

is sufficient to give a. Suppose $a = 1{\cdot}6$ is the trial value; calculate $f(a) - k$ for $a = 1{\cdot}5$, $a = 1{\cdot}55$, $a = 1{\cdot}60$, $a = 1{\cdot}65$ in succession, and obtain by interpolation a value of a which makes $f(a) = k$ as nearly as possible. Then test the resulting value against other parts of the record. Given a, A is easily found.

Another method, which perhaps uses the data more completely, is to use the equation,

Average income in the range x_1 to x_2

$$= \frac{a}{a-1}\left(\frac{1}{x_1{}^{a-1}} - \frac{1}{x_2{}^{a-1}}\right) \div \left(\frac{1}{x_1{}^a} - \frac{1}{x_2{}^a}\right).$$

For various workings on the formula see House of Commons Committee on the Income Tax (H. of C. No. 365 of 1906, pp. 220–30, 240–1, 245–6).

Makcham's Formula.

The equation $-\dfrac{1}{y} \cdot \dfrac{dy}{dx} = a + bc^x$ leads to a formula important in actuarial work.

For convenience in integration write

$$a = \frac{1}{\cdot} \log s, \quad b = - \log c \times \log g.$$

Then $\log y = x \log s + c^x . \log g + \text{const.}$

$\therefore y = ks^x . (g)^{c^x}$, where k, s, g, c are constants . . . . . (88)

This is Makeham's formula where y is the number of a given generation who survive to the age x, and is written l_x.

The ratio of the number of persons dying in an interval, δx, to the number alive at the beginning of the interval, divided by the duration of the interval, is

$$\frac{l_x - l_{x+\delta x}}{l_x . \delta x} = -\frac{1}{l_x} . \frac{dl_x}{dx} . \quad . \quad . \quad . \quad . \quad . \quad (89)$$

when the interval is indefinitely diminished. This expression is written μ_x and is called " the force of mortality."

The differential equation of the formula then gives

$$\mu_x = a + bc^x \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (90)$$

and the assumption is that the force of mortality is the sum of two quantities one of which is a constant a, and the other $bc^x = \mu'_x$, say, is such that it increases in a constant geometrical progression, for $\dfrac{D_x\mu'_x}{\mu'_x} = \log c.$

A more complicated form, obtainable by writing $\overline{a + a'x}$ for a above, is used by Mr. Hardy (*loc. cit.* p. 88) and in the *Report for* 1912–13 *on the Administration of the National Insurance Act,* Part I., p. 585 (Cd. 6907).

He also uses a hyperbolic equation for graduation,

$$z = \log \frac{y}{N - y} = k + \frac{m}{a + x} + \frac{n}{b + x},$$

where y is the number of husbands below age x and N is the total number of husbands (*Construction of Mortality Tables,* pp. 50–1 and Cd. 6907, p. 595).

CHAPTER VI.

THEORY OF CORRELATION.

Introductory.

ONE of the principal classes of problems in statistics is to determine whether phenomena are independent of each other, and if not, to measure their dependence.

In this chapter we consider principally the problem as it arises in connection with two or three variable quantities, the causes of whose variation may have something in common.

Suppose that we have pairs of observations, *e.g.* the height and span of a man, the heights of pairs of brothers, the income and rent of a household. Let the pairs of measurements be (X_1, Y_1) (X_2, Y_2), etc., and let there be a frequency group of the X's and another of the Y's, with averages $\bar{x}, \bar{y}$. Then if X and Y are completely independent, when we are told a value of X, we shall have no knowledge about the magnitude of the corresponding value of Y ; the chance that it shall have any particular deviation from $\bar{y}$ is simply that given by its own frequency curve ; but if there is anything common to X and Y in the causes of their variations, the statement of the value of an X will presumably affect the probability of the deviations of the corresponding Y.

X and Y may of course be connected rigidly by an equation, as, for example, X lbs. and Y kilos. may be different ways of expressing the weight of the same body, so that $X = 2 \cdot 204Y$, and Y/X is constant. In the cases with which we have to deal, however, the connection is not one of direct relation ; when X is given, Y is not determinate, but in a series of measurements (*e.g.* of height) we shall find for the same X varying values of Y.

If the average or shape of the frequency curve of the Y's associated with a given X is not the same as that for all values

350

of Y when the sorting by values of X is not made, then•there is something common to the two quantities, and they are said to be correlated. •

•An obvious first method of analysis is to arrange the observed values of Y in " arrays," each array containing those values for which the X is the same, as in an ordinary cross table. The average $\bar{Y}_t$ of an array of Y's when $X = X_t$ would, if there were complete independence and the number of observations were great, tend to equal $\bar{y}$, the average of all the Y's ; if it differs from $\bar{y}$ by more than would be expected in random sampling, then there is an indication that the value of Y is not independent of that of X.
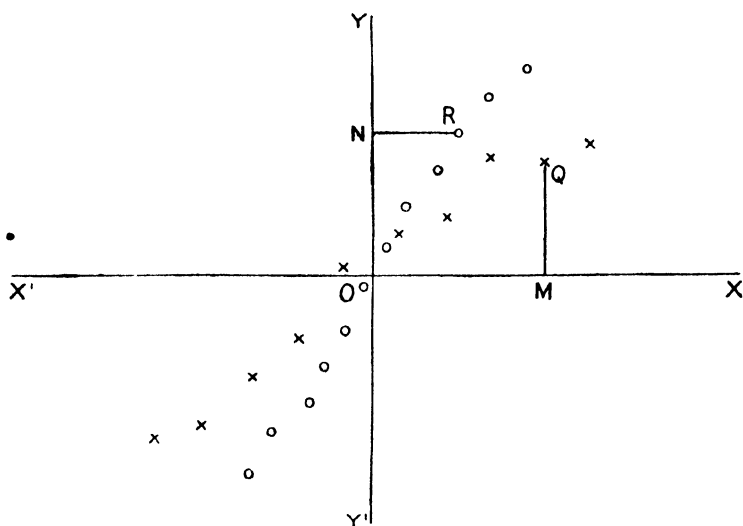
In the figure let O represent the averages of the X's and of the Y's. Let x_t, y_t, be the excess of X_t, Y_t over their averages, $\bar{x}, \bar{y}$, and let OM be a selected x_t and MQ be the average of the y's in that array, so that $MQ = \bar{Y}_t - \bar{y}$; and let the marks × × .. indicate various positions of Q.

Then if Y is independent of X, × × × will lie away from XX' only if the observations are not sufficiently numerous to give the true averages. If Y is not independent of X, Q will tend to have a definite locus, which a free-hand line drawn through its various positions will approximately define. If this is the case we can write $\bar{Y}_t - \bar{y} = f(X_t - \bar{x})$, so that when X_t

is given, though the *actual* value of Y_t is not known, yet $\bar{Y}_t$, the *average* of the array found in repeated selections, is approximately determinate.

Similarly, if we take an array of X's corresponding to a selected value of y_t (ON), the averages of these arrays, such as R, marked o o o, tend to lie on a curve $\bar{X}_t - \bar{x} = f_1 (Y_t - \bar{y})$, where f_1 is not the same as f.

The locus of Q is called the curve of " regression " of Y on X, and that of R the curve of regression of X on Y.

It frequently happens that these curves are approximately rectilinear, especially in the neighbourhood of O, so that $\bar{Y}_t - \bar{y} = kx_t$ approximately, where k is a constant when x_t is small.

The gradient of this line, k, equals $\dfrac{MQ}{OM}$ approximately, for any small value of OM, and may presumably be found by some method of averaging the various values of $\dfrac{MQ}{OM}$. We return to this on p. 355 and p. 364.

We can approach the problem from a different aspect as follows.

Let there be n magnitudes $X_1, X_2 \ldots$, and n magnitudes $Y_1, Y_2 \ldots$

Select at random an X, and *independently* select a Y, and form the product XY. Then in the long run, when a particular X happens to be selected, the various values of Y will come with equal frequency, and in the long run each of the n^2 products $X_1Y_1, X_1Y_2 \ldots X_2Y_1 \ldots X_nY_n$ will occur with the same frequency.

The sum of a very great number, N, of the products

$$= S(XY) = S(\bar{x} + x)(\bar{y} + y) = N\bar{x}\bar{y} + \bar{x} \cdot Sy + \bar{y}Sx + Sxy.$$

Here Sy tends to be $\dfrac{N}{n}(y_1 + y_2 + \ldots + y_n) = 0$, and Sx also tends to o.

$S(xy)$ tends to be $\dfrac{N}{n^2} (x_1y_1 + x_1y_2 + \ldots + x_2y_1 + \ldots + x_ny_n)$

$$= \dfrac{N}{n^2} \cdot Sx \cdot Sy = 0.$$

Hence S(XY) tends to $N\bar{x}\bar{y}$, and the mean of the product XY tends to equal the product of the means of X and of Y.

But if the selection of Y is *not* independent of that' of X, the n^2 products $x_1 y_1$ etc. do not come with equal frequency, and mean $XY = \bar{x}\bar{y} +$ Mean xy.

Again if $\bar{m}$ is the unweighted average of $M_1\, M_2 \ldots$, where in the notation of pp. 319, 320 $M_t = \bar{m} + m_t$, and $\bar{m}_w$ is the weighted average, where W_t the weight for $M_t, = \bar{w} + w_t$, $\bar{w}$ being the average of the weights.

$$\bar{m}_w = \frac{S(\bar{w} + w_t)(\bar{m} + m_t)}{n\bar{w}} = \frac{1}{n\bar{w}}(n\bar{w}\bar{m} + Sw_t m_t)$$

$$= \bar{m}\left(1 + \frac{1}{n}S\frac{w_t}{\bar{w}}\cdot\frac{m_t}{\bar{m}}\right) \quad \ldots \ldots \quad (91)$$

$\bar{m}_w = \bar{m}$ only if $Sw_t m_t = 0$, and this will only be the case if n is large, and if on the whole a large weight is not more often found with a large than with a small quantity, or *vice versa*.

On p. 288 it was shown that with the notation above, the mean of $(x + y)^2$ was $\sigma_x^2 + \sigma_y^2$, where σ_x and σ_y are the standard deviations of X and Y, if the selections of X and Y are quite independent.

It is easy to see that when there is dependence the analysis is modified and leads to

$$s^2 = \text{Mean } (x + y)^2 = \sigma_x^2 + \sigma_y^2 + 2\,.\,\text{Mean } xy \quad . \quad . \quad (92)$$

The Coefficient of Correlation.

Hence it appears that the quantity *Mean xy* enters into many expressions when X and Y are not independent and that in itself it gives an indication of the existence and amount of correlation. However, its magnitude depends on the units used in measuring x and y so that there is no natural scale for it, and consequently a quantity defined as follows is used in preference to it.

If X_1Y_1, X_2Y_2, $\ldots X_tY_t \ldots X_nY_n$ are pairs of measurements, and the averages and standard deviations of the X's and Y's are

A A*

$\bar{x}$, σ_x, $\bar{y}$, σ_y, then the *coefficient of correlation* between X, Y is written r_{xy}, where

$$r_{xy} = \frac{S\{(X_t - \bar{x})(Y_t - \bar{y})\}}{n\sigma_x\sigma_y} = \frac{1}{n}S\left(\frac{x_t}{\sigma_x} \cdot \frac{y_t}{\sigma_y}\right)$$

(taking $X_t = \bar{x} + x_t$, $Y_t = \bar{y} + y_t$) ₀

$$= \frac{1}{n\sigma_x\sigma_y}\{S(X_tY_t) - \bar{x}SY_t - \bar{y}SX_t + n\bar{x}\bar{y}\}$$

$$= \frac{1}{n\sigma_x\sigma_y}\{S(X_tY_t) - n\bar{x}\bar{y}\}, \quad . \quad . \quad . \quad . \quad . \quad . \quad (93)$$

since $SY_t = n\bar{y}$, $SX_t = n\bar{x}$.

In the examples just given

$$\text{Mean XY} = \bar{x}\bar{y} + r_{xy}\sigma_x\sigma_y$$

$$\overline{m_w} = \overline{m}\left(1 + r_{mw}\frac{\sigma_m}{\overline{m}} \cdot \frac{\sigma_w}{\overline{w}}\right)$$

$$s^2 = \sigma_x{}^2 + \sigma_y{}^2 + 2r_{xy}\sigma_x\sigma_y.$$

Write r for r_{xy}. It can readily be shown that r is never > 1 or < -1.

For $\quad\quad\quad\quad n^2r^2\sigma_x{}^2\sigma_y{}^2 = (Sx_ty_t)^2$;

but $\quad n^2\sigma_x{}^2\sigma_y{}^2 - (Sx_ty_t)^2$ ₒ

$$= (x_1{}^2 + x_2{}^2 + \ldots)(y_1{}^2 + y_2{}^2 + \ldots) - (x_1y_1 + x_2y_2 + \ldots)^2,$$

since $\quad n\sigma_x{}^2 = x_1{}^2 + x_2{}^2 + \ldots$, and $n\sigma_y{}^2 = y_1{}^2 + y_2{}^2 + \ldots$,

$$= (x_1y_2 - x_2y_1)^2$$

$$+ (x_1y_3 - x_3y_1)^2 + \ldots + (x_2y_3 - x_3y_2)^2 + \ldots + (x_ny_{n-1} - x_{n-1}y_n)^2$$

which is > 0, unless $x_1y_2 - x_2y_1 = 0 = x_1y_3 - x_3y_1 = \ldots$, and

$$\therefore \quad\quad \frac{x_1}{y_1} = \frac{x_2}{y_2} = \frac{x_3}{y_3} = \ldots = \frac{x_n}{y_n} = \pm\frac{\sigma_x}{\sigma_y},$$

in which case the expression $= 0$, and $r = \pm 1$.

$$\therefore \quad r^2 = \frac{(Sx_ty_t)^2}{n^2\sigma_x{}^2\sigma_y{}^2} < 1, \text{ and } 1 > r > -1, \text{ unless } y \text{ varies directly as } x,$$

and then $\quad\quad\quad\quad r = +1 \text{ or } -1. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (94)$

Hence r is a quantity which depends on all the observations, is zero when independence is complete and *Mean xy* $= 0$, is independent of the units in which X and Y are measured, increases whenever a positive x_t is found with a positive y_t or a negative x_t with a negative y_t, but only reaches the value $+1$ (which it can never exceed) when x and y are connected rigidly by the equation $y = x \times$ constant. If positive x's

are found with negative y's and *vice versa*, r varies from o to $-$ I.

r is therefore a sensitive measurement of the amount of correlation.

To distinguish it from other measurements it is sometimes called the sum-product coefficient of correlation.

If the pairs are grouped in arrays, such as $x_{s\,1}y_s$; $x_{s\,2}y_s$, ... and $\bar{y}_s$ is the average of the n_t quantities $_1y_s$, $_2y_s$ ..., then

$$Sxy = Sx_s \cdot n_s\bar{y}_s, \text{ and } \frac{\sigma_y}{\sigma_x}.r = S\left(\frac{n_s\,x_s^2}{n\sigma_x^2} \cdot \frac{\bar{y}_s}{x_s}\right), \text{ where } Sn_sx_s^2 = n\sigma_x^2$$

$\dfrac{\sigma_y}{\sigma_x}\,r$ is therefore a weighted average of the ratios $\dfrac{MQ}{OM}$ on p. 352,

and $y = r\,\dfrac{\sigma_y}{\sigma_x}\,.\,x$ is an approximation to the locus of Q.

On the following pages we examine the circumstances which give r various numerical values, study the distribution of X, Y on various hypotheses, and find the equations of the lines of regression.

Nature of r.

Let X and Y be two variable quantities which depend on other variables U, V, W in such a way that

$$X_t = {}_1U_t + {}_2U_t + \ldots + {}_pU_t + {}_1V_t + {}_2V_t + \ldots + {}_qV_t,$$
$$Y_t = {}_1U_t + {}_2U_t + \ldots + {}_pU_t + {}_1W_t + {}_2W_t + \ldots + {}_qW_t,$$

where $_1U_t$ is selected at random from a frequency group of any form whose mean is $_1\bar{u}$ and standard deviation $_1\sigma_u$, $_2U_t$ is selected independently from another group, and so on throughout the U's, V's and W's. p and q are any integers.

Write $_1U_t = {}_1\bar{u} + {}_1u_t$ etc., and $X_t = \bar{x} + x_t$, $Y_t = \bar{y} + y_t$ where $\bar{x}$ and $\bar{y}$ are the means of all possible values of X and Y. Let σ_x, σ_y be the standard deviations of X and Y.

Then $\quad x_t = {}_1u_t + {}_2u_t + \ldots + {}_pu_t + {}_1v_t + {}_2v_t + \ldots + {}_qv_t,$
$$y_t = {}_1u_t + {}_2u_t + \ldots + {}_pu_t + {}_1w_t + {}_2w_t + \ldots + {}_qw_t;$$
$$\sigma_x^2 = {}_1\sigma_u^2 + \ldots + {}_p\sigma_u^2 + {}_1\sigma_v^2 + \ldots + {}_q\sigma_v^2,$$
and $\quad \sigma_y^2 = {}_1\sigma_u^2 + \ldots + {}_p\sigma_u^2 + {}_1\sigma_w^2 + \ldots + {}_q\sigma_w^2,$

since, by hypothesis, the u's, v's, and w's are all independent of each other (p. 288).

If $_1\sigma_u = {}_2\sigma_u = \ldots = \sigma_u, {}_1\sigma_v = {}_2\sigma_v = \ldots = \sigma_v$, and $_1\sigma_w = {}_2\sigma_w = \ldots = \sigma_w$, or if σ_u^2, σ_v^2, σ_w^2 are mean values of the U, V, and W standard deviations squared, then

$$\sigma_x^2 = p\sigma_u^2 + q\sigma_v^2, \quad \sigma_y^2 = p\sigma_u^2 + q\sigma_w^2.$$

A A* 2

Also, mean $x_t y_t = $ mean $_1 u_t^2 + $ mean $_2 u_t^2 + \ldots$
$$+ \text{mean } _1u_t \cdot _1w_t + \ldots + \text{mean } _1v_t \cdot _1u_t + \ldots$$
$$= p\sigma_u^2,$$

for, since, by hypothesis, the selections of the various U's, V's and W's are independent of each other, such a term as mean $_1u_t \cdot _1w_t$ is zero in the long run.

Hence

$$r = \text{mean } x_t y_t / \sigma_x \cdot \sigma_y = \frac{p\sigma_u^2}{\sqrt{\{(p\sigma_u^2 + q\sigma_v^2)(p\sigma_u^2 + q\sigma_w^2)\}}}, \quad (95)$$

and, in particular, if $\sigma_u = \sigma_v = \sigma_w$,

$$r = \frac{p}{p+q} \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad (96)$$

This is the simplest conception of the numerical value of r; expressed in words it shows that the correlation coefficient tends to be the ratio of the number of causes common in the genesis of two variables to the whole number of independent causes on which each depends.

If constants a, b, c, d are introduced so that

$$\left. \begin{aligned} x_t &= a_1 \cdot _1u_t + \ldots + a_p \cdot _pu_t + b_1 \cdot _1v_t + \ldots + b_q \cdot _qv_t \\ y_t &= c_1 \cdot _1u_t + \ldots + c_p \cdot _pu_t + d_1 \cdot _1w_t + \ldots + d_q \cdot _qw_t \end{aligned} \right\} \cdot \quad (97)$$

then $\quad \sigma_x^2 = S \cdot a^2\sigma_u^2 + S \cdot b^2\sigma_v^2, \quad \sigma_y^2 = S \cdot c^2\sigma_u^2 + S \cdot d^2\sigma_w^2,$

$$\text{mean } x_t y_t = Sac\sigma_u^2,$$

and the expression for r can be readily written down.

The Correlation Surface.

Consider the case where the frequency curves of the U's, V's and W's are normal, and in the first place let us examine the grouping of X,Y in the simple case where

$$x_t = u_t + v_t, \ y_t = u_t + w_t.$$

The chance of the concurrence of selections u_t, v_t, w_t is

$$\frac{1}{\sigma_u\sqrt{2\pi}} e^{-\frac{u_t^2}{2\sigma_u^2}} \times \frac{1}{\sigma_v\sqrt{2\pi}} e^{-\frac{v_t^2}{2\sigma_v^2}} \times \frac{1}{\sigma_w\sqrt{2\pi}} e^{-\frac{w_t^2}{2\sigma_w^2}}.$$

Eliminate v_t and w_t.

The chance of obtaining x_t (to $x_t + \delta x$), y_t (to $y_t + \delta y$), when a particular value u_t (to $u_t + \delta u$) has been selected from the U group, is

$$\frac{1}{\sigma_u \sigma_v \sigma_w (2\pi)^{\frac{3}{2}}} e^{-\frac{1}{2}\left\{ \frac{u_t^2}{\sigma_u^2} + \frac{(u_t - x_t)^2}{\sigma_v^2} + \frac{(u_t - y_t)^2}{\sigma_w^2} \right\}} . \, \delta x \, \delta y \, \delta u$$

$$= \frac{1}{\sigma_u \sigma_v \sigma_w (2\pi)^{\frac{3}{2}}} . \, e^{-\frac{1}{2}k\{u_t - lx_t - my_t\}^2 - \frac{1}{2}(ax_t^2 + 2hx_t y_t + by_t^2)} . \, \delta x \, \delta y \, \delta u$$

where
$$k = \frac{1}{\sigma_u^2} + \frac{1}{\sigma_v^2} + \frac{1}{\sigma_w^2}, \ kl = \frac{1}{\sigma_v^2}, \ km = \frac{1}{\sigma_w^2}$$

$$a = \frac{1}{\sigma_v^2} - kl^2, \ b = \frac{1}{\sigma_w^2} - km^2, \ h = -klm.$$

The chance (say P_{xy}) of obtaining x_t, y_t from *any* value of u is obtained by adding the chances of obtaining them from an assigned value of u.

Hence, writing x, y for x_t, y_t,

$$P_{xy} = \frac{1}{\sigma_u \sigma_v \sigma_w . 2\pi} e^{-\frac{1}{2}(ax^2 + 2hxy + by^2)} . \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}k(u_t - lx_t - my_t)^2} . \, du$$

$$= \frac{1}{\sigma_u \sigma_v \sigma_w 2\pi \sqrt{k}} e^{-\frac{1}{2}(ax^2 + 2hxy + by^2)}$$

Now
$$\sigma_x^2 = \sigma_u^2 + \sigma_v^2, \ \sigma_y^2 = \sigma_u^2 + \sigma_w^2, \ r\sigma_x \sigma_y = \sigma_u^2$$

$$\therefore \ k = \frac{(\sigma_u^2 + \sigma_v^2)(\sigma_u^2 + \sigma_w^2) - \sigma_u^4}{\sigma_u^2 \sigma_v^2 \sigma_w^2} = \frac{\sigma_x^2 \sigma_y^2 (1 - r^2)}{\sigma_u^2 \sigma_v^2 \sigma_w^2}$$

$$a = kl - kl^2 = l(k - kl) = l\left(\frac{1}{\sigma_u^2} + \frac{1}{\sigma_w^2} \right) = \frac{1}{k\sigma_v^2} . \frac{\sigma_y^2}{\sigma_u^2 \sigma_w^2}$$

$$= \frac{1}{\sigma_x^2 (1 - r^2)}$$

and similarly

$$b = \frac{1}{\sigma_y^2 (1 - r^2)}$$

$$-h = \frac{1}{k} . kl . km = \frac{1}{k \sigma_v^2 \sigma_w^2} = \frac{\sigma_u^2}{\sigma_x^2 \sigma_y^2 (1 - r^2)} = \frac{r}{\sigma_x \sigma_y (1 - r^2)}$$

$$\therefore \ P_{xy} = \frac{1}{2\pi \sigma_x \sigma_y \sqrt{1 - r^2}} e^{-\frac{1}{2(1-r^2)}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - \frac{2rxy}{\sigma_x \sigma_y} \right)} \quad . \quad . \quad . \ (98)$$

By an extension of this method it is shown (Elderton, *Frequency Curves*, pp. 109 *seq.*, following Pearson, *Trans.*

of Royal Society, vol. 187 (1896), A, 175) that if x_t, y_t are formed by the weighted sum of a number of variables, all of normal frequency, as expressed in the equations (97) above, P_{xy} is of the form $\kappa e^{-(ax^2 + 2hxy + by^2)}$ as just found, and when the conditions that this surface shall have unit volume, standard deviations σ_x, σ_y and mean product $r\sigma_x\sigma_y$ are expressed by integration, the values of κ, a, h, b are the same as in the simple case discussed.

Though this method is of considerable interest, and by it the measurement of correlation by the product-sum formula was introduced into modern statistics by Professor Pearson, its importance is greatly diminished by the assumption that the elemental frequency curves are normal. The following analysis is free from this assumption ; it is derived from Professor Edgeworth's paper on *The Law of Great Numbers*, to which reference has already been made.

Edgeworth's Method.

Let

$$x_t = {}_1u_t + {}_2u_t + \ldots + {}_pu_t \ldots + {}_nu_t, \quad y_t = {}_1v_t + {}_2v_t + \ldots + {}_pv_t \ldots + {}_nv_t,$$

where ${}_1u_t$ is the deviation from its mean of a quantity selected from a curve of frequency whose standard deviation is ${}_1\sigma_u$, and ${}_2u_t \ldots {}_nu_t$, ${}_1v_t \ldots {}_nv_t$ have similar meanings. Let the selection of the various u_t's be quite independent of each other, so that mean ${}_1u_t \; {}_2u_t$ etc. tend to zero, and let the v_t's be similarly independent ; but let some, at any rate, of the v's be not independent of the u's, so that mean ${}_1u_t \cdot {}_1v_t$, mean ${}_2u_t \cdot {}_2v_t$, . . . mean ${}_nu_t \cdot {}_nv_t$ do not all tend to zero. Such a quantity as mean ${}_1u_t \cdot {}_2v_t$ is, however, to be taken as tending to zero.

Let n be large and $\dfrac{1}{\sqrt{n}}$ negligible, and the other conditions described above (p. 299) be satisfied, so that the curves of frequency of x and y taken separately are normal curves of error.

Then $\sigma_x{}^2 = S({}_p\sigma_u{}^2)$, $\sigma_y{}^2 = S({}_p\sigma_v{}^2)$,

and *mean xy* $= S(\text{mean } {}_pu_t \cdot {}_pv_t)$, ... (99)

where p is any integer from 1 to n.

Now rotate the axis on which x and y are measured

through an angle θ determined by $\tan 2\theta = \dfrac{2r\sigma_x\sigma_y}{\sigma_x{}^2 - \sigma_y{}^2}$, where $r = \dfrac{\text{mean } xy}{\sigma_x\sigma_y}$ is the coefficient of correlation between x and y. Thus we write

$$X_t = x_t \cos\theta + y_t \sin\theta, \quad Y_t = x_t \sin\theta - y_t \cos\theta.$$

Similarly write

$$_pU_t = {}_pu_t \cos\theta + {}_pv_t \sin\theta, \quad _pV_t = {}_pu_t \cdot \sin\theta - {}_pv_t \cos\theta.^{*}$$

Then $X_t = S{}_pU_t$, and $Y_t = S{}_pV_t$.

$\text{Mean } X_tY_t = \sin\theta\cos\theta\,(\sigma_x{}^2 - \sigma_y{}^2) - \cos 2\theta \cdot (\text{mean } xy) = 0$,
from the value assigned to $\tan 2\theta$.

$$\sigma_X{}^2 = \cos^2\theta \cdot \sigma_x{}^2 + \sin^2\theta \cdot \sigma_y{}^2 + \sin 2\theta \cdot r\sigma_x\sigma_y = S\,(\text{mean }{}_pU_t{}^2)$$

$$\sigma_Y{}^2 = \sin^2\theta \cdot \sigma_x{}^2 + \cos^2\theta \cdot \sigma_y{}^2 - \sin 2\theta \cdot r\sigma_x\sigma_y = S\,(\text{mean }{}_pV_t{}^2)$$

$$\therefore \ \sigma_X{}^2 + \sigma_Y{}^2 = \sigma_x{}^2 + \sigma_y{}^2$$

$$\sigma_X{}^2 - \sigma_Y{}^2 = (\sigma_x{}^2 - \sigma_y{}^2)\cos 2\theta + 2r\sigma_x\sigma_y \sin 2\theta = (\sigma_x{}^2 - \sigma_y{}^2)\sec 2\theta$$

$$= \sqrt{\{(\sigma_x{}^2 - \sigma_y{}^2)^2 + 4r^2\sigma_x{}^2\sigma_y{}^2\}}$$

$$4\sigma_X{}^2\sigma_Y{}^2 = 4\sigma_x{}^2\sigma_y{}^2\,(1 - r^2)$$

$\text{mean }{}_pU_t \cdot {}_pV_t = \sin\theta\cos\theta\,({}_p\sigma_u{}^2 - {}_p\sigma_v{}^2) - \cos 2\theta\,(\text{mean }{}_pu_t \cdot {}_pv_t)$.

$$\therefore \ S(\text{mean }{}_pU_t \cdot {}_pV_t) = \text{mean } X_tY_t = 0 \quad \ldots \ldots \quad (100)$$

Now follow the method of pp. 296–7 above for calculating the moments of X and Y.

Let a, β be any small constants, whose use is to collect similar terms.

$$e^{aX_t + \beta Y_t} = e^{a\,{}_1U_t + \beta\,{}_1V_t} \times e^{a\,{}_2U_t + \beta\,{}_2V_t} \times \ldots$$

Expand the exponentials, give t all possible values and take their mean, remembering that the mean of a sum is the sum of the means of its terms, and that the mean of a product of independent factors (as are the factors on the right-hand side) is the product of the means of the factors, and that the mean of first powers is zero.

In the expansion of the left-hand side, the coefficient of $a^k\beta^l$ occurs in mean $(aX + \beta Y)^{k+l} \div (k+l)!$, and equals mean $(X^kY^l) \div (k!\,l!)$, where k, l are any integers.

---

* These meanings of X, Y, U, V are of course not connected with the use of the same letters on p. 355 above.

The right-hand side = product of n factors
$$\{1 + \tfrac{1}{2}a^2 \,(\text{mean }_pU_t{}^2) + \tfrac{1}{2}\beta^2 \,(\text{mean }_pV_t{}^2) + a\beta \,(\text{mean }_pU_t{}_pV_t) + \ldots\},$$

$$\therefore \qquad \log\left(1 + \ldots + \frac{a^k\beta^l}{k!\,l!} \,(\text{mean } X^kY^l) + \ldots\right),$$

where k or l may be zero,
$$= S[\log\{1 + \tfrac{1}{2}a^2(\text{mean}_pU_t{}^2) + \tfrac{1}{2}\beta^2(\text{mean}_pV_t{}^2) + a\beta(\text{mean}_pU_t{}_pV_t) + \ldots\}]$$
$$= (\text{expanding by the logarithmic series and adding terms})$$
$$\tfrac{1}{2}a^2 S(\text{mean }_pU_t{}^2) + \tfrac{1}{2}\beta^2 S(\text{mean }_pV_t{}^2) + a\beta S(\text{mean }_pU_t{}_pV_t) + \ldots$$
$$= \tfrac{1}{2}a^2 \cdot \sigma_X{}^2 + \tfrac{1}{2}\beta^2 \cdot \sigma_Y{}^2 + a\beta \times 0^* + \text{terms involving } a^3, a^2\beta, \text{ etc.}$$

By arguments similar to those on p. 297 it is found that the terms in a^3 are of order $\dfrac{1}{\sqrt{n}}$ in comparison with terms in a^2, and hence, when $\dfrac{1}{\sqrt{n}}$ is neglected,

$$1 + \ldots + \frac{a^k\beta^l}{k!\,l!} \,(\text{mean } X^kY^l) + \ldots = e^{\frac{1}{2}a^2\sigma_X{}^2} \times e^{\frac{1}{2}\beta^2\sigma_Y{}^2}$$

$$= \left(1 + \tfrac{1}{2}a^2\sigma_X{}^2 + \ldots + \frac{1}{k!}(\tfrac{1}{2}a^2\sigma_X{}^2)^k \ldots\right)\left(1 + \tfrac{1}{2}\beta^2\sigma_Y{}^2 + \ldots + \frac{1}{l!}(\tfrac{1}{2}\beta^2\sigma_Y{}^2)^l \ldots\right).$$

Equate coefficients of terms on the two sides of this equation.

We have (when $l=0$) mean $X^{2k+1}=0$, mean $X^{2k}=\dfrac{(2k)!}{2^k k!}\sigma_X{}^{2k}$ as in the normal curve, and similarly for Y when $k = 0$.

All means involving odd powers of X or Y are zero.

$$\textbf{Mean }(X^{2k}Y^{2l}) = \frac{(2k)!}{2^k k!}\sigma_X{}^{2k} \cdot \frac{(2l)!}{2^l l!}\sigma_Y{}^{2l} \quad \ldots \ldots \ldots \text{(101)}$$

These are precisely the mean powers found by integrating the surface

$$z = \frac{1}{2\pi\sigma_X\sigma_Y}e^{-\frac{1}{2}\left(\frac{X^2}{\sigma_X{}^2}+\frac{Y^2}{\sigma_Y{}^2}\right)} = \frac{1}{\sigma_X\sqrt{2\pi}}e^{-\frac{1}{2}\frac{X^2}{\sigma_X{}^2}} \cdot \frac{1}{\sigma_Y\sqrt{2\pi}}e^{-\frac{1}{2}\frac{Y^2}{\sigma_Y{}^2}},$$

where X is independent of Y. Hence, as on pp. 297-8, we may take this equation as giving the frequency of X, Y.

It remains to transfer back to the original axes.

As already shown $\sigma_X\sigma_Y = \sigma_x\sigma_y\sqrt{(1 - r^2)}$
$$X^2\sigma_Y{}^2 + Y^2\sigma_X{}^2$$
$$= (x\cos\theta + y\sin\theta)^2\sigma_Y{}^2 + (x\sin\theta - y\cos\theta)^2\sigma_X{}^2$$
$$= x^2(\cos^2\theta\sigma_Y{}^2 + \sin^2\theta\sigma_X{}^2) + y^2(\sin^2\theta\sigma_Y{}^2 + \cos^2\theta\sigma_X{}^2) - xy\sin 2\theta(\sigma_X{}^2 - \sigma_Y{}^2)$$

---

* From equation (100).

Then sum of the coefficients of x^2 and y^2
$$= \sigma_X{}^2 + \sigma_Y{}^2 = \sigma_x{}^2 + \sigma_y{}^2,$$
and their difference $= \cos 2\theta(\sigma_Y{}^2 - \sigma_X{}^2) = \sigma_y{}^2 - \sigma_x{}^2,$
hence the coefficients are $\sigma_y{}^2$ and $\sigma_x{}^2$.

Also $\sin 2\theta(\sigma_X{}^2 - \sigma_Y{}^2) = \tan 2\theta(\sigma_x{}^2 - \sigma_y{}^2) = 2r\sigma_x\sigma_y.$

Hence $\dfrac{X^2}{\sigma_X{}^2} + \dfrac{Y^2}{\sigma_Y{}^2} = \dfrac{1}{\sigma_x{}^2\sigma_y{}^2(1 - r^2)}\{x^2\sigma_y{}^2 + y^2\sigma_x{}^2 - 2r\sigma_x\sigma_y \cdot xy\}$

and the equation of the surface is

$$z = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{(1 - r^2)}} \cdot e^{-\frac{1}{2(1-r^2)}\left\{\frac{x^2}{\sigma_x{}^2} + \frac{y^2}{\sigma_y{}^2} - \frac{2rxy}{\sigma_x\sigma_y}\right\}} \qquad . \quad . \ (102)$$

as already found (formula (98)) on the simpler hypothesis that the elemental groups have normal frequency.

In this equation $\sigma_x{}^2 = S\sigma_u{}^2$, $\sigma_y{}^2 = S\sigma_v{}^2$, $r\sigma_x\sigma_y = S \cdot (r_p \cdot {}_p\sigma_u \cdot {}_p\sigma_v),$ from equations (99), where r_p is the correlation coefficient between ${}_pu$ and ${}_pv$.

Constants can be introduced in the original equations so that $\quad x = {}_1a_1u + {}_2a_2u + \ldots$ and $y = {}_1b_1v + {}_2b_2v + \ldots$ without affecting the method of analysis.

Properties of the Normal Correlation Surface.

The centre is at the average of the x and of the y variables.

Volume $= \iint z\,dx\,dy$

$$= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1 - r^2}} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2(1-r^2)}\left(\frac{x}{\sigma_x} - r\frac{y}{\sigma_y}\right)^2} \cdot e^{-\frac{1}{2}\frac{y^2}{\sigma_y{}^2}} \cdot dx\,dy = 1$$

$$\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \ (103)$$

Second moment in $y = \iint z y^2 dx\,dy$

$$= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1 - r^2}} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2(1-r^2)} \cdot \frac{x'^2}{\sigma_x{}^2}} \cdot y^2 \cdot e^{-\frac{1}{2}\frac{y^2}{\sigma_y{}^2}} \cdot dx' \cdot dy,$$

where $x' = x - r\dfrac{\sigma_x}{\sigma_y} \cdot y$; then integrating in respect of x' we find that the expression

$$= \frac{1}{\sigma_y\sqrt{2\pi}} \int_{-\infty}^{\infty} y^2 e^{-\frac{1}{2}\frac{y^2}{\sigma_y{}^2}} dy = \sigma_y{}^2 \text{ by formula (21).} \quad . \quad . \quad . \ (104)$$

and similarly the second moment in x is $\sigma_x{}^2$.

Mean product of $xy = \iint zxy dx dy$ $\pm \infty$ being the limits of integration,

$$= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1 - r^2}} \iint \left(x' + r\frac{\sigma_x}{\sigma_y}y\right) e^{-\frac{1}{2(1 - r^2)} \cdot \frac{x'^2}{\sigma_x^2}} \cdot ye^{-\frac{1}{2}\frac{y^2}{\sigma_y^2}} dx' dy$$

$$= \frac{1}{\sigma_y\sqrt{2\pi}} \int r\frac{\sigma_x}{\sigma_y}y^2 e^{-\frac{1}{2}\frac{y^2}{\sigma_y^2}} dy = r\sigma_x\sigma_y \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (105)$$

$$\iint zyx^2 dx dy = 3r\sigma_x^2\sigma_y, \quad \iint zx^2y^2 dx dy = (2r^2 + 1)\sigma_x^2\sigma_y^2 \quad . \quad . \quad . \quad (106)$$

The section by every plane parallel to XOZ, YOZ is a normal curve. *E.g.*, if $x = x_1$,

$$z = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1 - r^2}} \cdot e^{-\frac{1}{2(1 - r^2)\sigma_y^2}\left(y - r\frac{\sigma_y}{\sigma_x}x_1\right)^2} \cdot e^{-\frac{1}{2}\frac{x_1^2}{\sigma_x^2}} \quad . \quad . \quad . \quad (107)$$

which is a normal curve with its centre at $y = r\dfrac{\sigma_y}{\sigma_x}x_1$, standard deviation $\sigma_y\sqrt{1 - r^2}$, and maximum ordinate

$$\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1 - r^2}} e^{-\frac{1}{2}\frac{x_1^2}{\sigma_x^2}}.$$

The frequency group of the y's corresponding to an assigned value of x is therefore normal, and its standard deviation is independent of x. The average of the group (and its mode and median) is for all values of x on the line $y = r\dfrac{\sigma_y}{\sigma_x}x = x \tan \phi_1$, (say), and this is the line of regression (p. 352).

$r\dfrac{\sigma_y}{\sigma_x}$ is the coefficient of regression of y in relation to x.

Similarly for a given value of y_1 the frequency group in x is normal, its standard deviation is $\sigma_x\sqrt{1 - r^2}$, and its average is on the line $x = r\dfrac{\sigma_x}{\sigma_y}y$, say $y = x \tan \phi_2$.

$r\dfrac{\sigma_x}{\sigma_y}$ is the coefficient of regression of x in relation to y.

The geometric mean of the two coefficients of regression is r.

Horizontal sections are similar ellipses. Thus if $z = z_1$,

$$\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - 2r\frac{xy}{\sigma_x\sigma_y} = -2(1 - r^2)\log(z_1 2\pi\sigma_x\sigma_y\sqrt{1 - r^2}) \quad . \quad . \quad (108)$$

The major axis of any such ellipse (where we take $\sigma_x > \sigma_y$) makes the angle θ with the axis of x, where

$$\tan 2\theta = \frac{2r\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2}.$$

Now $\tan 2\phi_1 = \frac{2r\sigma_x\sigma_y}{\sigma_x^2 - r^2\sigma_y^2}$, and $\tan 2\phi_2 = \frac{2r\sigma_x\sigma_y}{r^2\sigma_x^2 - \sigma_y^2}$.

If $r = \pm 1$, $\theta = \phi_1 = \phi_2$, and the surface degrades into the plane $\frac{y}{\sigma_y} = \frac{x}{\sigma_x}$. Otherwise, when $|r| < 1$, $\phi_2 > \theta > \phi_1$, and the lines of regression lie on either side of the plane of the major axis of the ellipses, and on either side of $\frac{y}{\sigma_y} = \frac{x}{\sigma_x}$. If $\sigma_x = \sigma_y$, $\theta = \frac{\pi}{4}$, and the lines of regression are equally inclined to the planes containing the principal axes of the ellipses.

It should be noticed that the surface is completely determined by five quantities, viz., two averages, two standard deviations, and one correlation coefficient.

Rectilinear Regression.

We have found that under certain conditions, of a simple nature and dependent mainly on plurality of causation, the line of regression, that is the locus of the averages of one variable (y) for given values of the other (x), is straight and passes through the position representing the general averages.

If the conditions are not rigidly, but only approximately, satisfied, there is a presumption that rectilinearity of regression will be approximately attained.

It may well happen that regression is still approximately rectilinear even if the variables x and y are not normally distributed, and that the equation $y = r\frac{\sigma_y}{\sigma_x}x$ may still be the equation of regression, though the surface of distribution is no longer determinable from the value of r.

Let there be n_t values of y in the array corresponding to a value x_t of x, and let their average be $\bar{y}_t$. Write $m_t = \frac{\bar{y}_t}{x_t}$, so that m_t is

the gradient of the line of regression determined from one group only.

Then it is shown above (p. 355) that $\tan \phi_1, = r\dfrac{\sigma_y}{\sigma_x}$, is a weighted average of $m_1, m_2 \ldots m_s \ldots$, the weights being $n_s x_s^2$ &c.
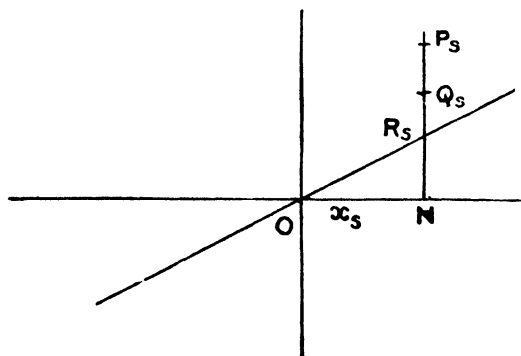
Let $ON = x_s$, NP_s be any value of y found with x_s, and $NQ_s = \bar{y}_s$ be the average of n_s such values. Let a line $y = ax + b$ meet NP_s at R_s.

Then Mr. Yule shows (*Statistical Journal*, 1897, pp. 817–8) that the sum of all values of $(R_sP_s)^2$, *i.e.*, $S[y_s - (ax_s + b)]^2$, extended over all values of x_s, is least when $b = 0$ and $a = r\dfrac{\sigma_y}{\sigma_x}$. This method depends on "least squares," for which see Appendix, Note 10.

This line $y = r\dfrac{\sigma_y}{\sigma_x}x$ then passes through the observations in such a way that the sum of the squares of the distances of the points representing the observations, measured from it parallel to the axis of y, is a minimum. This line is, then, whatever the distribution, a good single representation of regression.

We can proceed a step further, if we assume that the dispersion of the y's in any s^{th} array is independent of the value of x_s, and is always σ_2. For if the averages tend to lie on a straight line, and only fail to do so exactly because of the paucity of observations, then the deviation from the average, namely R_sQ_s, has a curve of frequency $Ke^{-\dfrac{(R_sQ_s)^2}{2\sigma^2}}$ where $\sigma^2 = \dfrac{\sigma_2^2}{n_s}$ (p. 312).

Hence the joint probability of deviations R_1Q_1, R_sQ_s ... is

$K'\epsilon^{-\frac{1}{2\sigma^2}Sn_s(R_sQ_s)^2}$, and this is greatest when $Sn_s(R_sQ_s)^2$ is least.

$Sn_s(R_sQ_s)^2 = S\{n_s(\bar{y}_s - (ax_s + b))^2\}$

$\qquad = S(n_s\bar{y}_s^2) + a^2S(n_sx_s^2)$

$\qquad\qquad + Nb^2 - 2aS(n_s\bar{y}_sx_s) - 2bSn_s\bar{y}_s + 2abSn_sx_s,$

where $\qquad N = n_1 + \ldots + n_s + \ldots.$

Here $Sn_s\bar{y}_s = 0 = Sn_sx_s$ since the origin is the double average.

$$S(n_sx_s^2) = Sx^2 = N\sigma_x^2, \quad S(n_s\bar{y}_sx_s) = Sxy = Nr\sigma_x\sigma_y$$

and the expression equals

$$Sn_s\bar{y}_s^2 + N(a\sigma_x - r\sigma_y)^2 + Nb^2 - Nr^2\sigma_y^2.$$

This is least when $b = 0$, and $a = r\dfrac{\sigma_y}{\sigma_x}$ as before.

The line $y = r\dfrac{\sigma_y}{\sigma_x}x$ is then the most probable locus of regression, if we assume rectilinearity and independence between deviation in an array and the corresponding value of x, the deviations from rectilinearity being due to fewness of observations.

The value of $Sn_s(R_sQ_s)^2$ reckoned from this line is $Sn_s\bar{y}_s^2 - Nr^2\sigma_y^2$.

If, however, there is nothing in the genesis of the measurements, or in their results, to justify the assumption of rectilinearity, r ceases to be an intelligible measurement of the *amount* or degree of commonness of causation, though it may still be a useful function of the quantities in analysis.

The Correlation Ratio.

To obtain a measurement completely independent of assumptions about distribution of the observations, Professor Pearson has devised the correlation ratio (*Drapers' Company Research Memoirs*, Biometric Series, II., 1905).

Let $_s\sigma_y$ be the standard deviation of the sth array, so that $n_s \cdot {_s\sigma_y}^2 = S(y_s - \bar{y}_s)^2$, and write $\cdot\sigma_a^2$ for the weighted mean of $_1\sigma_y^2$, $_2\sigma_y^2$ ... so that $N\sigma_a^2 = S(n_s \cdot {_s\sigma_y}^2) = SS(y_s - \bar{y}_s)^2$, the inner summation being extended over an array, and the outer indicating the sum over all the arrays.

Write $N\sigma_m^2 = S(n_s\bar{y}_s^2)$, so that σ_m^2 is the weighted mean square of the averages of the arrays.

Then $N\sigma_y^2 = S(y^2)$, the summation being extended over all values of y, and

$$N\sigma_a^2 = S\{S(y_s^2) - 2\bar{y}_s S y_s + n_s\bar{y}_s^2\} = S(S y_s^2 - n_s\bar{y}_s^2) = S(y^2) - S(n_s\bar{y}_s^2).$$

$\therefore \sigma_y^2 = \sigma_m^2 + \sigma_a^2$, as is otherwise evident.

Now write
$$\eta = \frac{\sigma_m}{\sigma_y} = \sqrt{\left(1 - \frac{\sigma_a^2}{\sigma_y^2}\right)} \quad \cdots \cdots \quad (109)$$

η is then called the *correlation ratio*. It is the ratio of the scattering of the averages of the arrays to the scattering of the group not regimented into arrays.

$\eta = 0$ only if $\sigma_m = 0$, and therefore if every $\bar{y}_s = 0$; that is, if the average of every array is coincident with the general average of the group.

$\eta = 1$ only if $\sigma_a = 0$, that is if every $_s\sigma_y = 0$, and the terms in each array are concentrated at a single point, $\bar{y}_s$.

Otherwise $1 > \eta > 0$.

In normal correlation every $_s\sigma_y = \sigma_y\sqrt{(1 - r^2)}$ formula (107); and then $\sigma_a^2 = \sigma_y^2(1 - r^2)$, and $\eta^2 = r^2$.

In other cases we have, as shown above, p. 365,

$$S n_s(R_s Q_s)^2 = N\sigma_m^2 - N r^2 \sigma_y^2 = N(\eta^2 - r^2)\sigma_y^2$$

$\therefore$
$$\eta^2 = r^2 + \frac{S n_s(R_s Q_s)^2}{N\sigma_y^2}, \quad \cdots \cdots \quad (110)$$

and $|\eta| > |r|$, unless every $R_s Q_s$ is 0 and the means of the arrays all lie on the line $y = r\dfrac{\sigma_y}{\sigma_x}x$.

We may now sum up the treatment of correlation so far.

If (x,y) is a pair of measurements (from their averages) of two variables (related in space, in time, in a thing or in an organism), and if when x is given as positive (or negative) there is a presumption that y is positive (or negative), or a presumption that y is negative (or positive), then the variables are said to be correlated. In such a case $\dfrac{1}{n}Sxy$ does not tend to zero when n is increased, but to a limit written as $r\sigma_x\sigma_y$. $r = 0$, $= 1$, $= -1$ have definite meanings; r is sensitive to all kinds of relationship between x and y. In general it

may be expected to be the greater as σ_a (the mean scattering within the arrays) is less.   If x and y are each the sum of $p + q$ independent elements of which p (only) are common to x and y, then r equals $p/(p + q)$, if the standard deviations of the elements are equal.   If x and y are generated linearly from a multiplicity of independent causes (some of them common to x and y), then r defines the whole frequency distribution of the pairs, the regression loci are rectilinear, and their equations are $y = r\dfrac{\sigma_y}{\sigma_x}.x$, and $x = r\dfrac{\sigma_x}{\sigma_y}.y$.   If the normal frequency surface cannot be assumed, but regression is rectilinear, the same equation is a good empirical state-ment of regression.   If nothing can be postulated as to the distribution of x and y or the averages of the arrays, the meaning of the numerical value of r is undefined (as is always the case with η when it is not o or 1).   In general, however, r may be said to measure the amount that is common in the systems of causation of x and y.

Correlation between Ungraded Variables.

The measurement of correlation by the methods so far discussed is only possible if we have adequate detailed observa-tions.   Cases of great interest arise when such detail is not forthcoming.

COLOUR OF HAIR.

Parent.

Son.	Light.	Dark.	Totals.
Dark . . .	a	b	n_1
Light . . .	c	d	n_2
Totals . .	m_1	m_2	N

Suppose that sons and parents are separated according to the colour of their hair distinguished as light or dark ;   and that of m_1 sons of light-haired parents c have light hair and the remaining a dark ; while of m_2 sons of dark-haired parents, d have light hair and the remaining b dark.   Let $a + b = n_1$, $c + d = n_2$, and $n_1 + n_2 = N = m_1 + m_2$.

Required to determine from these data whether there is

a relationship between hair-colour of sons and parents, and, if so, to measure it.

If in such a case normal distribution of the variable (say, amount of pigment) and normal correlation can be assumed, the problem is determinate. For the ratios $m_2 : N$, and $n_1 : N$ give (by inverse use of the table, p. 271) the abscissæ on the scales of pigment which correspond to the division between light and dark ; for any given value of r the fraction of the correlation surface bounded by planes through these abscissæ is known, and the equation of the fraction b/N to this is, conversely, an equation for r.

The necessary analysis is given by Pearson (*Phil. Trans.* A, Vol. CXCV, pp. 1 *seq.*) and Elderton (*Frequency Curves,* Chapter VII) and results in a troublesome equation for r which can be solved approximately.

If we have control of the data and can make both separations at the median, a simple solution can be given.

Suppose that intelligence in arithmetic and in algebra is normally distributed. Arrange a large class of N boys in order of intelligence (as known by marks or otherwise) in arithmetic ; now mark also their order in respect of algebra, and suppose that b are found above the median in both respects, c below in both, d above in arithmetic and below in algebra, and a above in algebra and below in arithmetic. It is not assumed that intelligence is measured, but only that an order can be assigned.

Then $a + b = \dfrac{N}{2} = c + d = a + c = b + d,$

$$\therefore \quad a = d = (\tfrac{1}{4} - q)N, \text{ say, and } b = c = (\tfrac{1}{4} + q)N.$$

It can be shown as follows that $r = \sin 2\pi q$.

Take the standard deviation on each scale to be unity.

Let the required surface be $\dfrac{1}{2\pi \sqrt{1 - r^2}}\, e^{-\frac{1}{2(1 - r^2)}(x^2 + y^2 - 2rxy)}$

The principal axis of the surface then makes $\dfrac{\pi}{4}$ with the axis

of x. $(\tfrac{1}{4} + q)$ equals the volume in the doubly-positive quadrant bounded by the planes $y = 0$, $x = 0$, and these planes cut off each of the similar elliptic horizontal sections $(\tfrac{1}{4} + q)$ of their area. Take the ellipse $x^2 + y^2 - 2rxy = 1$.
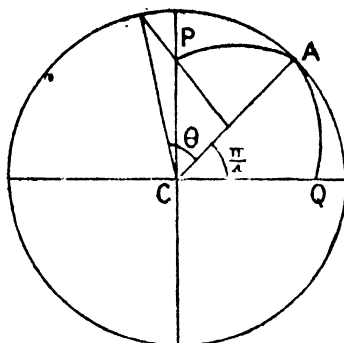
Hence in the figure

Elliptic area CPQ $= (\frac{1}{4} + q)$ of area of ellipse.

Let θ be the eccentric angle of P.
The ellipse referred to its principal axes is

$$(1 - r) x^2 + (1 + r) y^2 = 1.$$

$$\frac{\tan \theta}{\tan \frac{\pi}{4}} = \frac{\text{major axis}}{\text{minor axis}} = \sqrt{\left(\frac{1 + r}{1 - r}\right)}.$$

$$\therefore \qquad r = - \cos 2\theta.$$

$$\frac{1}{4} + q = \frac{2 \text{ area CPA}}{\text{area of ellipse}} = \frac{2\theta}{2\pi}.$$

$$\therefore \qquad 2\pi q = 2\theta - \frac{\pi}{2}$$

and $$\sin 2\pi q = - \cos 2\theta = r.$$

E.g., if 40 % of the boys were found to be above the median in both

$$\frac{1}{4} + q = \cdot 4, \quad q = \cdot 15, \quad r = \sin \frac{3}{10} \pi = \cdot 81.$$

If $q = 0$, $r = 0$; if $q = \frac{1}{4}$, $r = 1$.

In the table relating to 83 boys given by Mr. W. Brown (*Biometrika*, Vol. VII., p. 366), 11 boys are above the median in algebra, but below in arithmetic. Here $q = \frac{1}{4} - \frac{11}{83} = \cdot 12$, and $r = \cdot 68$. Mr. Brown using the complete order (and not merely the median) obtains $\cdot 65$, and using the marks obtains $\cdot 79$. All need correction, given by Mr. Brown, for age and position in school.   •

If normality of distribution of the two attributes cannot be postulated, the problem of measurement of the *amount* of correlation becomes indeterminate, and a number of methods have been tried.

Association.

The expected number of dark-haired fathers with dark-haired sons, if there were no causal connection, would be $\frac{m_2}{N} \times \frac{n_1}{N} \times N = \beta$, where out of N cases n_1 sons and m_2 parents were dark-haired.

The notation of p. 367 being adopted, and a, γ, δ being the number in the a, c, d compartments that would be the most probable in a chance arrangement, we have

$$a + b = a + \beta = n_1, \quad a + c = a + \gamma = m_1 \text{ etc.}$$

and
$$a - a = b - \beta = c - \gamma = \delta - d = b - \frac{m_2 n_1}{N}$$

$$= \frac{b(a + b + c + d) - (b + d)(a + b)}{N} = \frac{bc - ad}{N} = qN, \text{ say.}$$

Then q is a measure of association, but no definite meaning can be given to it except in extreme cases.

Instead of q, Mr. Yule takes $Q = \dfrac{bc - ad}{bc + ad}$ (the " coefficient of association ") or $\omega = \dfrac{\sqrt{bc} - \sqrt{ad}}{\sqrt{bc} + \sqrt{ad}}$ (the " coefficient of colligation ") as measurements. (See *Introduction to Theory of Statistics*, p. 37, and *Statistical Journal*, 1912, p. 593.) $Q = \omega = 0$, if $bc = ad$, and $q = 0$, the case of no association ; and $Q = \omega = 1$ if a or d is zero, and -1 if b or c is zero, which cases correspond to the maximum of association on this method.

These coefficients have therefore definite meanings in extreme cases, but the meaning of (*e.g.*) $Q = \frac{2}{3}$ can only be appreciated by the examination of numerous instances, and in the end it can hardly be affirmed that a greater Q means a greater amount of " association," for no definite measurable meaning has been given to the term " association."

Contingency.

If, instead of trying to find the *amount* of association, we ask for evidence of its *existence*, that is whether the observations could arise if the attributes were independent, we are on surer ground.

If $p : 1 - p$ is the ratio of dark to light-haired among sons in general, then from the observations $p = \dfrac{n_1}{N}$ is the best value we can assign.

Hence the chance that, if N sons were divided arbitrarily (*e.g.* according as their Christian names began with A to K or L to Z) into two groups containing respectively m_1 and m_2 of them, a would be found in the first group is that discussed above (p. 282-4), and may be written

$$\frac{1}{\sigma\sqrt{2\pi}}\,e^{-\frac{x^2}{2\sigma^2}},\ \left(\frac{1}{\sqrt{m_1}}\text{ being neglected}\right),$$

where $x = a - pm_1 = a - \dfrac{n_1 m_1}{N} = a - a = qN$ *

$$\sigma^2 = p(1 - p)\,m_1\left(1 - \frac{m_1}{N}\right) = \frac{n_1}{N}\cdot\frac{n_2}{N}\,m_1\cdot\frac{m_2}{N}.$$

The chance that so great a deviation, positive or negative, as $(a - a)$ should occur is

$$2\int_z^\infty \frac{1}{\sqrt{2\pi}}\,e^{-\frac{1}{2}z^2}\,dz,\text{ where } z = \frac{x}{\sigma}.$$

Notice that

$$\frac{x^2}{\sigma^2} = \frac{q^2 N^5}{n_1 n_2 m_1 m_2} = q^2 N^3\cdot\frac{(n_1 + n_2)\,(m_1 + m_2)}{n_1 n_2 m_1 m_2}$$

$$= q^2 N^3\left(\frac{1}{n_1 m_1} + \frac{1}{n_1 m_2} + \frac{1}{n_2 m_1} + \frac{1}{n_2 m_2}\right)$$

$$= \frac{(a - a)^2}{a} + \frac{(b - \beta)^2}{\beta} + \frac{(c - \gamma)^2}{\gamma} + \frac{(d - \delta)^2}{\delta} = X^2,\text{ say.}$$

E.g. in the distribution $\dfrac{65\ \ |\ \ 235}{35\ \ |\ \ 165}$

$$n_1 = 300,\quad n_2 = 200,\quad m_1 = 100,\quad m_2 = 400,\quad N = 500,$$
$$\sigma^2 = 19\cdot2,\quad a = 60,\quad x = qN = 5,$$

$\dfrac{x}{\sigma} = 1\cdot14,\quad F(1\cdot14) = \cdot373\text{ (p. 271)},\quad\text{and}\quad 2\{\tfrac{1}{2} - F(1\cdot14)\} = \cdot254.$

---

* q has the same meaning as in the previous paragraph and is not $1 - p$ as in Chapter II.

The chance against obtaining .65 or more or 55 or less, when 100 are selected out of 500, and the chance in one selection is 300 : 500, is ·746 to ·254 or about 3 to 1.

Given n_1, m_1, N and a, the remaining numbers b, c, d, n_2, m_2 are known, and the chance just found is equally the chance affecting any one of the numbers b, c, d taken independently of each other. It should not be spoken of as the chance of the distribution as a whole ; to find this we should need to know the chance p from a wider universe, and not as $\frac{n_1}{N}$ determined from a limited number of observations, and also the general chance to which $\frac{m_1}{N}$ is the approximation.

To illustrate this difficulty, we will consider a problem that has often been discussed.

	Not vaccinated.	Vaccinated.	Totals.
Recovered . .	a	b	n_1
Died. . . .	c	d	n_2
Totals . .	m_1	m_2	N

In an epidemic of smallpox the number of cases is N, of whom m_2 were vaccinated, n_2 died, and other categories are as shown in the table.

The recovery rate, as shown by the whole statistics, is $\frac{n_1}{N}$, and if vaccination (whether directly, or by the other attributes correlated with it) had nothing to do with recovery, then the chance of a vaccinated or an unvaccinated patient's recovery would be $\frac{n_1}{N}$, and the chance that as many as b recover out of m_2 vaccinated is

$$\int_{b-\beta}^{\infty} \frac{N!}{\sqrt{(2\pi n_1 n_2 m_1 m_2)}} \cdot e^{-\frac{1}{2} \cdot \frac{x^2 N^3}{n_1 n_2 m_1 m_2}} \cdot dx,$$

where $qN = b - \beta = x$; if this is small there is evidence of a relation between vaccination (or the circumstances that lead to it) and recovery.

The rate $\frac{n_1}{N}$, however, is subject to a standard deviation of

$\sqrt{\frac{n_1 n_2}{N^3}}$, and unless this is negligible its effect on the computed result should be tested.

A measure of the advantage (or disadvantage) of vaccination (apart from evidence of the existence of some effect) could conceivably be obtained by comparing the recovery rates $\frac{b}{m_2}$ and $\frac{a}{m_1}$, but apart from the statement of these rates and their standard deviations there is no direct method of procedure.

The question of the existence and of the measurement of association becomes more complicated when, instead of simple alternatives in each attribute, we have several different classes, for example several grades of hair colour both in father and son.

Professor Pearson has introduced the " coefficient of contingency " for the measurement of such a case (*Drapers' Company Research Memoir*, Biometric Series I, 1904 ; Elderton Chapter X).

Numbers of Observations
Classes of First Attribute.

		Classes of First Attribute				
	a_1	a_2	.	.	.	n_1
	b_1	b_2	.	.	.	n_2
	c_1	c_2	.	.	.	n_3
	.	.		.	.	. .
	.	.		.	.	. .
	m_1	m_2	.	.	.	N

Let $n_1; n_2 \ldots, m_1, m_2 \ldots$ be the totals of lines and columns as in the table, and N the total number of observations.

Then if n_1/N, m_1/N are assumed to be the accurate proportions of the first class in each attribute to the total, the most probable number to be found in the position of a_1, if there was no association, would be $a_1 = \frac{n_1}{N} \times \frac{m_1}{N} \times N$. Similarly values $a_2 \ldots, \beta_1, \beta_2 \ldots, \gamma_1, \gamma_2 \ldots$ can be computed for the other places.

The divergences $a_1 - a_1$, $a_2 - a_2$, $\ldots b_1 - \beta_1 \ldots$ afford some measure of association. Since an excess or defect is equally probable, it is convenient to take the squares, $(a_1 - a_1)^2$ etc., instead of the linear quantities.

Analogy with the case of four categories suggests the formation of the function

$$X^2 = \frac{(a_1 - a_1)^2}{a_1} + \frac{(a_2 - a_2)^2}{a_2} + \ldots + \frac{(b_1 - \beta_1)^2}{\beta_1} + \ldots + \ldots \quad (111)$$

This function also has a place in the measurement of the appropriateness of a formula to represent given observations (formula (130)).

The *coefficient of contingency* is then defined as

$$C = \frac{1}{\sqrt{1 + \dfrac{N}{X^2}}} \quad \ldots \quad \ldots \quad (112)$$

When $X = 0$ and there is no association, $C = 0$.

As X increases, C increases from 0, and tends towards 1 as $\dfrac{X^2}{N}$ becomes great.

$\dfrac{X^2}{N} = \dfrac{(a_1 - a_1)^2}{n_1 m_1} + \dfrac{(a_2 - a_2)^2}{n_1 m_2} + \ldots$, and depends only on ratios, not on the whole number measured.

It can be shown (Elderton, p. 147) that if the numbers $a_1, a_2 \ldots b_1$ etc. are those which occur in appropriate divisions of a normal correlation surface, and if the number of divisions is large, while N is so large that the smallest of the entries is not less than a small integer, then C approximates to r, the coefficient of correlation. This relation appears to have suggested the form of the function of X^2 which defines C.

The value obtained for C differs according to the number of divisions taken, and this consideration diminishes its utility as a measurement ; but a method has been given by which this difficulty can be overcome (*Biometrika*, Vol. IX, pp. 116–139).

The significance of particular values of C can only be appreciated by experience of many cases.

It should be noticed that C, and the analysis of the previous paragraph, can be applied to cases where classes can be defined, but have no measurable attributes.

Correlation of Time Series.

So far we have been concerned with the correlation between two statistical groups where the measurements all relate to the same time ; we have still to consider how to test the relation between two series, where the pairs $x_1, y_1 \ldots x_p, y_p \ldots$ x_t, y_t are measurements of quantities at successive intervals. Here it is generally the case that one value of x is not independent of those that come before or after it, and the relationship found between x and y may merely reflect a general or

periodic progress in time and not any more intimate connection. Nearly all series in time have a trend, and these trends, whether equally rapid or not, or in the same direction or not, will yield a high correlation coefficient even if the quantities are otherwise independent. For example the coefficient between

$$1,\ 2,\ 3\quad.\quad.\quad.\quad.\quad.\quad20$$
$$100,\ 98,\ 96\quad.\quad.\quad.\quad.\quad.\quad62$$

where $x_t = t$ and $y_t = 100 - 2(t - 1)$, is -1.

We need to find the correlation between the deviations after the time element is eliminated.

One method * is to obtain smoothed lines for each quantity, as above (pp. 132 *seq.*), to compute the differences between the observations of each year and the values given by the smoothed line, and to treat these differences as the quantities whose correlation should be measured ; *i.e.* to measure the correlation between such quantities as those represented on the diagram (facing p. 155).

If the series are markedly periodic, the result would be only to bring out the correlation due to the periodicities, and these are better studied by harmonic analysis. And if the series are strongly " compensated " (p. 148), so that a positive deviation is generally followed by a negative one, the correlation would reflect this symptom.

But if the oscillations are random, so that apart from a regular trend the measurement of one year is unrelated to the measurement of adjacent years, the coefficient, calculated between the deviations from the smoothed lines, measures the same kind of relation as that already discussed in the correlation of groups.

Let x_p, y_p be a pair of measurements in the pth year, and let $\bar{x}_p$, $\bar{y}_p$ be the averages of the measurements of m years of which the pth is central, m being odd. Then the correlation coefficient to be calculated is that between $x_p - \bar{x}_p$ and $y_p - \bar{y}_p$.

The method can also be applied if the smoothing is effected by the method recommended by Professor Persons (*The Review of Economic Statistics*, No. 1, 1919, Harvard University Press). In this method the average $\bar{y}$ is calculated for m years during

* See Hooker, *Statistical Journal*, 1901, pp. 485 *seq.*

which the trend appears to be in one direction and without any marked change of gradient, and the smoothed line is assumed to be of the form $y - \bar{y} = kt$, where t is the number of years from the centre of the period. k is determined by the condition that the sum of the squares of the deviations from this line shall be a minimum, viz., that $S\{y_t - (\bar{y} + kt)\}^2$ is a minimum, where y_t is the observation t years from the centre. Then $k = \dfrac{St(y_t - \bar{y})}{St^2}$.

If $m = 2n + 1$, $St^2 = 2(1^2 + \ldots + n^2) = \dfrac{n(n+1)(2n+1)}{3}$.

This method overcomes partially a difficulty that occurs when moving averages are used in a case where the trend is continually concave (or convex), and the averages always below (or above) the observations.

Another method, introduced by Miss F. E. Cave (Royal Society's *Proceedings*, Vol. LXXIV, p. 407, 1904) and by Mr. Hooker (*Statistical Journal*, 1905, pp. 696 *seq.*) and more recently developed by Professor Karl Pearson, Miss B. M. Cave and others, is to correlate not the observations but the differences between successive observations. A period of $m + 1$ years is selected, in which the observations are $x_0, x_1 \ldots x_m$ and $y_0, y_1 \ldots y_m$, and the coefficient of correlation between the pairs $x_1 - x_0$, $y_1 - y_0$, $x_2 - x_1$, $y_2 - y_1 \ldots$ $x_m - x_{m-1}$, $y_m - y_{m-1}$ is calculated.

Since the average of the quantities $x_1 - x_0$, $x_2 - x_1 \ldots$ is $\dfrac{1}{m}(x_m - x_0)$, the deviations from the average which alone enter into the formula for r are the excess (or defect) of the increment in a particular year over the mean increment, or they may be described as the annual variations from the trend. If the smoothed lines of x and y are markedly concave or convex the correlation will be dominated by this symptom, but if the observations oscillate in an irregular way about a straight line, we shall obtain a measure of correlation independent of the element of time.

To get over the difficulty arising from concavity or convexity a more elaborate method, named by Professor Pearson that of " variate difference correlation," has been devised.*

---

* *Biometrika*, Vol. X, pp. 179 *seq.* and pp. 340 *seq.*

This is based on the assumption that x_p can be expressed as $x_p = X_p + bt_p + ct_p{}^2 + \ldots$, where X_p is independent of the influence of time, and the effect of the time can be expressed as a parabolic function: and similarly $y_p = Y_p + b't_p + \ldots$

$$x_{p+1} - x_p = X_{p+1} - X_p + b + c\,(2t_p + 1) + \ldots$$

Mr. Hooker's method ignores c and further constants.

The second difference gives

$$x_{p+1} - 2x_p + x_{p-1} = X_{p+1} - 2X_p + X_{p-1} + 2c + 6dt_p \ldots$$

and its use ignores d etc., assuming a strict parabolic form.

It can be shown that when time is eliminated the correlation between any differences equals the correlation between X_p and Y_p. The process is complete when the correlation coefficient is no longer affected by proceeding to a further difference.

There is, however, a great difficulty in applying this method to any differences except perhaps the first three, owing to the want of precision or the small number of significant figures in ordinary observations. The effect can be seen if we take the squares of the numbers 2·6, 2·7 . . . when written only to the first decimal place.

	Δ	Δ²
6·8		
7·3	·5	0
7·8	·5	·1
8·4	·6	0
9·0	·6	0
9·6	·6	0
10·2	·6	·1
10·9	·7	

The second differences if written completely are all ·02.

The method is, in fact, too refined for ordinary statistical observations.

The difference between the methods may be exhibited as follows.

The x quantity that is correlated, if we use the fourth difference, is $6\{x_0 + \tfrac{1}{6}(x_2 + x_{-2}) - \tfrac{2}{3}(x_1 + x_{-1})\}$, where the suffixes mark the distance to right or left of the centre; here the extreme terms increase the expression.

If we take the moving average based on five terms, the quantity is

$$x_0 - \tfrac{1}{5}(x_{-2} + x_{-1} + x_0 + x_1 + x_2) = \tfrac{4}{5}(x_0 - \tfrac{1}{4}(x_2 + x_{-2}) - \tfrac{1}{4}(x_1 + x_{-1}))$$

and the extreme terms diminish the expression.

The eighth difference gives

$$x_0 - \tfrac{1}{2}(x_1 + x_{-1}) + \tfrac{2}{7}(x_2 + x_{-2}) - \tfrac{4}{35}(x_3 + x_{-3}) + \tfrac{1}{70}(x_4 + x_{-4})$$

while the moving average gives

$$x_0 - \tfrac{1}{8}(x_1 + x_{-1}) - \tfrac{1}{8}(x_2 + x_{-2}) - \tfrac{1}{8}(x_3 + x_{-3}) - \tfrac{1}{8}(x_4 + x_{-4}).$$

On the other hand, the second difference gives

$$x_0 - \tfrac{1}{3}(x_1 + x_0 + x_{-1}),$$

and is the *same* as that obtained from a moving average based on only three terms, and therefore subject to a very considerable chance error.

The various methods need further examination and more experience of their applicability. It appears that the moving average does not give the right importance to extreme terms, while the difference method is too sensitive to the effect of roughness in observations. In either case, the resulting measurement of correlation depends on the assumptions made, and is not so easily intelligible as in the measurement of correlation of groups.

Graphic Comparison of Series.

Apart from the determination of a measurement of correlation, the problem arises of how best to exhibit the relationship graphically.

The following method is useful. Let $x_1, x_2 \ldots x_n, y_1, y_2 \ldots y_n$ be deviations from moving averages (as in the table on p. 387), or (if there is no trend) be actual measurements, and let $\bar{x}$ and $\bar{y}$ be their averages.

Make a graph of the values of y on any convenient scale, time being measured horizontally. Then the x values may be placed on this diagram on any scale and with any origin.

Take b as the origin for x, and let 1 unit of x correspond to c units of y. b and c are to be chosen. A convenient method is to select them so that the sum of the squares of the vertical distances between the points representing pairs such as x_1, y_1 shall be a minimum (Appendix, Note 10).

That is $S\{c(x + b) - y\}^2$ is to be a minimum.

By differentiating with regard to b and to c, it is found that $c = \dfrac{S(x - \bar{x})(y - \bar{y})}{n\sigma_1^2}$, and $c(\bar{x} + b) = \bar{y}$, where σ_1 is the standard deviation of the x's.

The averages of the deviations should therefore be marked at the same point on the vertical scale, and the differences from their average of the x's should be multiplied by $\dfrac{r\sigma_2}{\sigma_1}$ and then measured on the y scale above and below the average, σ_2 being the standard deviation of the y's and r the coefficient of correlation.

An example is given in measuring unemployment in the *Statistical Journal*, 1912, pp. 799–800.

Note added in* 1936.—Professor R. M. Fréchet [*] has developed an interesting relationship between r and η (the correlation ratio).

With the notation of pp. 363–6, $N\eta^2\sigma_y^2 = S(n_s\bar{y}_s^2)$

$$N r \sigma_x \sigma_y' = S(xy) = S(n_s \bar{y}_s . x_s)$$

$$\therefore r^2 = \eta^2 \times \frac{(Sn_s\bar{y}_s x_s)^2}{S(n_s\bar{y}_s^2)\,S(x^2)}, \text{ or } r = \eta \times \rho.$$

Here ρ is the coefficient of correlation when in each array all the observations are collected at the average, that is n_s objects are at x_s, $\bar{y}_s$, etc., for $N\sigma_x^2 = S(x^2) = S(n_s x_s^2)$.

Thus r can be resolved into two factors : the correlation ratio and another correlation coefficient. " Nous avons fait observer que ce facteur ρ est celui qui dépend seulement de la *forme* de la relation tandis que la *rigueur de la dépendance* n'interirent que dans η." " Deux facteurs l'un, ρ, qui ne dépend que de la forme de la ligne des moyennes, l'autre, η, qui n'en dépend pas ou à peu près pas."

Notice that if $\eta = 1$, $r = \rho$, while (as shown on p. 366 above) every $\sigma_y = 0$. In such a case, when an x is given the value of y is completely determined.

If $\rho = 1$, the line of means is perfectly rectilinear.

Maximum Value of the Coefficient of Contingency.—If relationship is perfect, all the objects in a given class of the first attribute are found in one class of the second. Without loss of generality (in a table where there are l rows and l columns) we can write this condition as $a_1 = n_1 = m_1$, $b_1 = n_2 = m_2 \ldots$, while all other compartments are empty.

Then $\alpha_1 = n_1^2/N$, $\beta_2 = n_2^2/N \ldots$, and $0 = \alpha_2 = \alpha_3 \ldots = \beta_1 = \beta_3 \ldots$

Now $X^2 = a_1^2/\alpha_1 - 2a_1 + \alpha_1 +$ similar terms

$$= \Sigma(a_1^2/\alpha_1) - 2N + N = \Sigma a_1^2/\alpha_1 - N$$

$$= n_1^2 \div n_1^2/N + n_2^2 \div n_2^2/N + \text{to } l \text{ terms} - N = (l-1)N.$$

$\therefore C = 1 \div \sqrt{(1 + 1/(l-1))} = \sqrt{(1 - 1/l)}$, which tends to unity as l is increased, but is definitely less than unity for ordinary small values of l.

---

[*] See *Revue de l'Institut International de Statistique*, 3 Annee, Livraison 4, pp. 365–79, especially pp. 366–7.

CHAPTER VII.

EXAMPLES OF CORRELATION.

In this section the results of several experiments and observations are given to illustrate the theory discussed above, and to show the arithmetical working of the measurements.

It should be premised that the theoretical value of r would only be obtained exactly in an infinite number of observations. It is shown in the chapter on probable errors that r, as calculated from n pairs, may differ from its true value by an amount whose standard deviation measured on the normal scale of error is $\dfrac{1 - r^2}{\sqrt{n}}$. Thus in the first example the correlation coefficient is known to be $\cdot 6$; 24 pairs are taken, and we should expect to be within $\dfrac{1 - \cdot 6^2}{\sqrt{24}} = \cdot 13$ of $\cdot 6$, while it is very unlikely that the difference would amount to 3 times $\cdot 13$. Conversely, if we do not know the coefficient *a priori*, we must read with our calculated value $\pm \dfrac{1 - r^2}{\sqrt{n}}$.

Some of the examples are intended to show simply the arithmetical methods of working out r from the observations.

In others when the observations are numerous the averages of arrays are obtained and comparison is made with the equation $y - \bar{y} = r \dfrac{\sigma_2}{\sigma_1} (x - \bar{x})$, which is the locus of these averages if regression is rectilinear.

In the final example the distribution of 1,000 pairs is compared in detail with the distribution given by the theoretical correlation surface.

In general x and y are measured not from their averages but from an arbitrary origin and then $r = \left(\dfrac{\mathrm{S}xy}{n} - \bar{x}\bar{y} \right) \div \sigma_1 \sigma_2,$ by formula (93).

380

Example 1.—To obtain a simple illustration of correlation when all the circumstances were known and the coefficient could be stated *a priori*, digits were taken from a mathematical table at random. x_t was taken as the sum of 5 digits, and y_t also as the sum of 5 digits of which 3 were included in the 5 which made x_t and 2 were different, and 24 pairs $(x_1y_1) \ldots (x_ty_t) \ldots$ were formed. The correlation coefficient for such pairs is $\frac{3}{5}$ (formula (96)). In the example in which only 24 pairs were taken it was ·537 ; the standard deviation of the coefficient $\frac{3}{5}$ is $\dfrac{1 - ·36}{\sqrt{24}} = ·13$, so that the deficit from so small a number is not remarkable.

The following table shows the working.

x	y	x^2	y^2	xy	
22	32	484	1,024	704	
27	27	729	729	729	
12	19	144	361	228	
21	30	441	900	630	
21	26	441	676	546	
27	26	729	676	702	
23	25	529	625	575	
17	22	289	484	374	
25	23	625	529	575	$n = 24$ $\bar{x} = 23\frac{1}{12}$ $\bar{y} = 23\frac{1}{3}$
11	9	121	81	99	$24\sigma_1{}^2 = 13706 - 24\,(23\frac{1}{12})^2$
16	24	256	576	384	$\sigma_1 = 6·18$
20	28	400	784	560	$\sigma_2 = 5·36$
37	29	1,369	841	1,073	
33	25	1,089	625	825	$r = \dfrac{13354 - 24\bar{x}\bar{y}}{24\sigma_1\sigma_2}$
18	20	324	400	360	
24	26	576	676	624	$= ·537$
22	17	484	289	374	
17	16	289	256	272	
32	27	1,024	729	864	
29	29	841	841	841	
26	17	676	289	442	
27	20	729	400	540	
26	26	676	676	676	
21	17	441	289	357	
554	560	13,706	13,756	13,354	

The arithmetic is simpler if x and y are measured from an origin 23 in each case.

Example 2.—Where we have few and sporadic observations, it is simpler to work out the arithmetic in full. For example the infantile mortality in 26 towns is in the adjoining table compared with the population (to the nearest 1,000) of these towns. r is only twice its standard deviation, and its exact value is therefore uncertain, but there is evidence that the

larger .the towns the higher the mortality. To attack the question of the causes of infantile mortality seriously, it would of course be necessary to take many more instances ard to consider many other factors besides crude population.

POPULATION AND INFANTILE MORTALITY IN 26 TOWNS.

Population.	Mortality.	
x	y	xy
000.		
55	162	8,910
39	201	7,839
36	241	8,676
35	162	5,670
31	179	5,549
30	174	5,220
27	176	4,752
24	208	4,992
24	163	3,912
23	206	4,738
22	172	3,784
20	200	4,000
19	218	4,142
19	198	3,762
19	132	2,508
16	155	2,480
15	148	2,220
15	220	3,300
15	141	2,115
12	169	2,028
7	155	1,085
6	129	774
6	167	1,002
5	150	750
5	171	855
4	161	644

$$r = \frac{Sxy - n\bar{x}\bar{y}}{26\sigma_1\sigma_2}, \text{ where}$$

$$n = 26$$
$$\bar{x} = 20\cdot346$$
$$\bar{y} = 175\cdot31$$

$$= \cdot34$$

Standard deviation of $r = \dfrac{1 - \cdot34^2}{\sqrt{26}} = \cdot17$

Totals . 529   4,558   95,707

Averages 20·346   175·31   —

Also $\sigma_1 = 12\cdot2$ $\sigma_2 = 27\cdot9$

Example 3.—A good illustrative example of method is obtained from statistics arising from the North Sea Fisheries Investigation of the size of herrings in relation to the rings which appear on their bodies and which are believed to show their age, one ring being formed each year.

The averages of arrays lie very near the theoretic straight line of regression, in spite of the skewness of the original curves.

In the table the size is measured on the axis of y, with origin at 31 cm. and unit 1 cm., and the number of rings is measured on the axis of x with origin at 7 rings and unit 1 ring, and the numbers of cases are entered in a square table.

n_2 is the total of cases for a given value of y, and n_2y and n_2y^2 and their sums are obtained in the last two columns, which lead to the average and standard deviations of the rings. Similarly n_1 is the total of cases in an x array, and the sums of n_1x and n_1x^2 lead to the average and standard deviation of the sizes.

In the last line the average in each array is given, obtained in each x array by multiplying the numbers of cases by the corresponding values of y.

Underneath each number of cases is given in brackets the corresponding value of $x \times y$; thus in the column under $x = -1$ in the row $y = 3$, we have four cases and $xy = -3$, so that the contribution of these four cases to the sum of xy is $4 \times -3 = -12$. The various terms thus contributed are shown below grouped in the four quadrants.

The origins are so chosen as to include as many zero terms in Sxy as possible. (Compare Yule, *Theory of Statistics*, p. 183.)

HERRING. NUMBER OF RINGS AND SIZE (LENGTH IN CENTIMETRES).

Number of rings . x	4 -3	5 -2	6 -1	7 0	8 1	9 2	10 3	11 4	12 5	13 6	Totals. n_2	n_2y	n_2y^2
Size. cm. 35 4	—	—	1 (-4)	—	1 (4)	2 (8)	2 (12)	—	—	—	6	24	96
34 3	—	1 (-6)	4 (-3)	4	15 (3)	14 (6)	7 (9)	3 (12)	1 (15)	1 (18)	50	150	450
33 2	1 (-6)	1 (-4)	11 (-2)	26	26 (2)	22 (4)	11 (6)	3 (8)	3 (10)	1 (12)	105	210	420
32 1	1 (-3)	24 (-2)	49 (-1)	53	26 (1)	7 (2)	5 (3)	1 (4)	—	—	166	166	166
31 0	—	28	43	45	21	6	2	—	—	—	115	0	0
30 -1	1 (3)	15 (2)	21 (1)	16	7 (-1)	1 (-2)	—	—	—	—	61	-61	61
29 -2	2 (6)	3 (4)	5 (2)	—	—	—	—	—	—	—	10	-20	40
28 -3	1 (9)	1 (6)	3 (3)	2	—	—	—	—	—	—	7	-21	63
27 -4	—	3 (8)	—	—	—	—	—	—	—	—	3	-12	48
26 -5	—	1 (10)	—	—	—	—	—	—	—	—	1	-5	25
Totals n_1	6	77	137	146	96	52	27	7	4	2	554	431	1,369
n_1x	-18	-154	-137	0	96	104	81	28	20	12	$Sn_1x =$ 32		
n_1x^2	54	308	137	0	96	208	243	112	100	72	$Sn_1x^2 =$ 1330		
Averages in arrays .	30·17	30·85	31·34	31·65	32·25	32·92	33·07	33·3	—	—			

$\bar{x} = 32 + 554 = ·0578$. Average, 7·0578 rings.
$\sigma_1{}^2 = 1330 \div 554 - ·0578^2 - ·083^*$ $\sigma_1 = 1·521$.
$\bar{y} = 431 \div 554 = ·778$. Average, 31·778 cm.
$\sigma_2{}^2 = 1369 \div 554 - ·778^2 - ·083^*$ $\sigma_2 = 1·335$.

* Sheppard's corrections, Appendix, Note 5.

Sxy	++	--	+-	-+
4	52	3	7	4
16	88	30	2	6
24	66	21	—	12
45	24	12	9	6
84	30	12		4
63	12	10		22
36	26	9		3
15	14	6		48
18	15	9		49
4		24		
		10		
636	146			154

$$Sxy = 636 + 146 - 9 - 154 = 619$$
$$r = \frac{Sxy - 55\overline{x}\,\overline{y}}{554\,\sigma_1\sigma_2} = ·528$$

Standard deviation of $r = ·035$

$$\frac{\text{Length} - 31·778 \text{ cm.}}{1·335 \text{ cm.}} = ·528 \quad \frac{\text{Number of rings} - 7·0578}{1·521}$$

Number of rings.	Length deduced from equation. cm.	Average of arrays. cm.
4	30·36	30·17
5	30·82	30·85
6	31·29	31·34
7	31·75	31·65
8	32·21	32·25
9	32·68	32·92
10	33·14	33·07
11	33·60	33·3

Example 4.—The following example is given to illustrate the value that may be obtained for *r*, when in the nature of the case ·there can be little correlation. For *x* the last digit of each of the (7 figure) logarithms of the numbers 2500–2549 and 2600–2649 was taken ; for *y* the last digit of the logarithm of a number 50 greater than *x*, *i.e.* 2550–2599 and 2650–2699. *r* is found to be ·086, which is less than its standard deviation for 100 pairs.

At the same time is shown an alternative method of setting out the arithmetic, which in some cases is simpler than the other methods used in this section.

This method leads readily also to the calculation of the correlation ratio.

OCCURRENCES OF PAIRS OF DIGITS.

x	0	1	2	3	4	5	6	7	8	9	n_s	Sy	xSy	$n_s\bar{y}_s^2$
0	—	—	1	—	1	1	1	—	—	4	8	53	0	351
1	3	—	—	1	1	—	—	1	1	1	8	31	31	120
2	—	2	1	—	3	1	—	—	—	1	8	30	60	112
3	2	—	—	1	—	1	—	2	2	3	11	65	195	384
4	1	2	2	—	—	2	1	3	1	—	12	51	204	217
5	1	1	—	—	1	1	4	1	—	—	9	41	205	187
6	1	1	1	2	—	1	—	2	1	—	9	36	216	144
7	3	3	2	1	1	3	1	—	3	1	18	68	476	257
8	—	2	—	3	2	1	—	—	2	1	11	49	302	218
9	—	—	—	1	—	—	—	1	1	3	6	45	405	338
	11	11	7	9	9	11	7	10	11	14	100	469	2,184	2,328

$$\bar{x} = 4\cdot72 \quad \sigma_x = 2\cdot69 \quad \bar{y} = 4\cdot69 \quad \sigma_y = 3\cdot03 \quad n = 100$$

$$S(x - \bar{x})(y - \bar{y}) = Sxy - 100\bar{x}\bar{y} = S(xSy) - 100\bar{x}\bar{y} = 2184 - 2214 = -30$$

$$r = \frac{-30}{100\sigma_x\sigma_y} = -\cdot037. \quad \text{Standard deviation of } r \text{ is } \cdot1, \text{ approx.}$$

Here n_x is the number of times the various x digits 0, 1 . . . were found. Sy is the sum of the corresponding y's; *e.g.* in the first line we have $2+4+5+6+9\times4=53$. $\bar{y}_x$ is the average y for a given x, and equals $Sy \div n_x$;

and $\qquad\qquad n_x\bar{y}_x{}^2 = (Sy)^2 \div n_x.$

The correlation *ratio* is found from the last column (see p. 366).

$$100\sigma_m{}^2 = Sn_x(\bar{y}_x - \bar{y})^2 \ast = Sn_x\bar{y}_x{}^2 - 2\bar{y}Sn_x\bar{y}_x + n\bar{y}^2$$
$$= Sn_x\bar{y}_x{}^2 - n\bar{y}^2 = 2328 - 2200 = 128.$$

$\eta = \dfrac{\sigma_m}{\sigma_y} = \dfrac{1\cdot13}{3\cdot03} = \cdot37$, and has a considerable value though the correlation coefficient is insignificant.

Example 5.—The following table gives data from " The Report on Heights and Weights of New York City Children " for 3,405 boys aged 14–15.

Height.	Number.	Average weight as in report	—	—	Average weight from equation.	—
Origin 61 inches.		Origin 100 lbs.			Origin 100 lbs.	
x	n_x	$\bar{y}_x$	$n_x\bar{y}_x$	$x \cdot n_x\bar{y}_x$	—	$n_x\bar{y}_x{}^2$
— 12	1	— 12	— 12	144	— 49·6	144
— 9	1	— 20	— 20	180	— 36·6	400
— 7	13	— 18	— 234	1,638	— 27·9	4,212
— 6	59	— 19	— 1,121	6,726	— 23·6	21,299
— 5	96	— 17	— 1,632	8,160	— 19·2	27,744
— 4	190	— 14	— 2,660	10,640	— 14·9	37,240
— 3	283	— 12	— 3,396	10,188	— 10·5	40,752
— 2	349	— 8	— 2,792	5,584	— 6·2	22,336
— 1	440	— 3	— 1,320	1,320	— 1·9	3,960
0	434	+ 2	+ 868	0	+ 2·5	1,736
+ 1	400	+ 7	+2,800	2,800	+ 6·8	19,600
+ 2	355	+11	+3,905	7,810	+11·2	42,955
+ 3	307	+17	+5,219	15,657	+15·5	88,723
+ 4	200	+20	+4,000	16,000	+20·0	80,000
+ 5	137	+24	+3,288	16,440	+24·2	78,912
+ 6	78	+30	+2,340	14,040	+28·6	70,200
+ 7	34	+35	+1,190	8,330	+32·9	41,650
+ 8	15	+34	+ 510	4,080	+37·2	17,340
+ 9	6	+42	+ 252	2,268	+41·6	10,584
+10	7	+42	+ 294	2,940	+45·9	12,348
Totals .	3,405	—	11,479	134,945	—	622,135

* On p. 366 the y's are measured from their average. Here it is necessary to subtract $\bar{y}$ throughout.

C C*

$\bar{x} = \cdot 2270$ $\sigma_1 = 2\cdot99$. Average $61\cdot227$ inches.
$\bar{y} = 3\cdot371$ $\sigma_2 = 16\cdot3$.* Average $103\cdot37$ lbs.
$$r = \frac{Sxy - n\bar{x}\bar{y}}{n\sigma_1\sigma_2} = \left(\frac{134945}{3405} - \cdot7652\right) \div \sigma_1\sigma_2 = \cdot797$$

The regression equation is $\dfrac{\text{Weight} - 103\cdot37}{16\cdot3} = \cdot797$ of $\dfrac{\text{Height} - 61\cdot227}{2\cdot99}$,

or Weight $= 103.4 + 4\cdot345$ (Height $- 61.23$).

The weights obtained from this equation are given in the sixth column of the table and should be compared with average weights corresponding to various heights given in the third column. The agreement is close from about 57 inches to 70 inches ; but below 57 inches actual weights do not fall off so rapidly as in the formula. The regression is not in fact linear for low statures.

$$3405\sigma_m{}^2 = Sn_e\bar{y}_e{}^2 - 3405\bar{y}^2, \text{ and } \sigma_m = 13\cdot1$$
$$\eta = \sigma_m \div \sigma_2 = \cdot81$$

Here the correlation ratio is practically the same as the correlation coefficient.

Example 6.—The methods discussed on pp. 374–8 for measuring the correlation between two time-series are worked out by comparing the value of imports into the United Kingdom per head of the population with the marriage rate in England and Wales, year by year.

x is the excess of the imports in any year over the average of the five years of which the year in question is central. y is similarly obtained from the marriage rate.

$\bar{x} = - \cdot62, \bar{y} = - \cdot3, \sigma_1 = 36\cdot9, \sigma_2 = 3\cdot59, Sxy = 4309, n = 50$

$r = \dfrac{4309 \div 50 - \cdot3 \text{ of } \cdot62}{36\cdot9 \times 3\cdot59} = \cdot65$ with standard deviation $\dfrac{1 - \cdot65^2}{\sqrt{50}} = \cdot09$

This is the measurement of the correlation by the use of moving averages.

---

* Calculated from data not reproduced here.

Years.	Imports per head.			Marriage Rate, England and Wales.			xy	
	Annual.	5 years' average.	Deviation x	Annual.	5 years' average.	Deviation y	+	−
1845	£3·30	—	—	17·2	—	—	—	—
1846	3·15	—	—	17·2	—	—	—	—
1847	3·21	3·22	− 1	15·8	16·5	−7	7	—
1848	2·91	3·35	−44	15·9	16·5	−6	264	—
1849	3·52	3·55	− 3	16·2	16·5	−3	9	—
1850	3·97	3·78	+19	17·2	16·9	+3	57	—
1851	4·14	4·30	−16	17·2	17·2	0	0	—
1852	4·35	4·70	−35	17·4	17·4	0	0	—
1853	5·51	4·93	+58	17·9	17·2	+7	406	—
1854	5·51	5·34	+17	17·2	17·1	+1	17	—
1855	5·16	5·80	−64	16·2	16·9	−7	448	—
1856	6·16	5·86	+30	16·7	16·5	+2	60	—
1857	6·66	6·01	+65	16·5	16·5	0	0	—
1858	5·80	6·44	−64	16·0	16·7	−7	448	—
1859	6·26	6·71	−45	17·0	16·6	+4	—	180
1860	7·32	6·92	+40	17·1	16·5	+6	240	—
1861	7·50	7·45	+ 5	16·3	16·7	−4	—	20
1862	7·72	8·05	−33	16·1	16·7	−6	198	—
1863	8·45	8·40	+ 5	16·8	16·8	0	0	—
1864	9·26	8·86	+40	17·2	17·0	+2	80	—
1865	9·06	9·12	− 6	17·5	17·1	+4	—	24
1866	9·80	9·35	+45	17·5	17·0	+5	225	—
1867	9·05	9·41	−36	16·5	16·7	−2	72	—
1868	9·60	9·54	+ 6	16·1	16·4	−3	—	18
1869	9·54	9·68	−14	15·9	16·3	−4	56	—
1870	9·70	10·09	−39	16·1	16·4	−3	117	—
1871	10·49	10·48	+ 1	16·7	16·7	0	0	—
1872	11·13	10·85	+28	17·4	17·0	+4	112	—
1873	11·54	11·19	+35	17·6	17·1	+5	175	—
1874	11·39	11·35	+ 4	17·0	17·0	0	0	—
1875	11·39	11·47	− 8	16·7	16·7	0	0	—
1876	11·30	11·34	− 4	16·5	16·2	+3	—	12
1877	11·75	11·18	+57	15·7	15·7	0	0	—
1878	10·87	11·28	−41	15·2	15·3	−1	41	—
1879	10·59	11·29	−70	14·4	15·1	−7	490	—
1880	11·88	11·29	+59	14·9	15·0	−1	—	59
1881	11·37	11·52	−15	15·1	15·1	0	0	—
1882	11·73	11·59	+14	15·5	15·2	+3	42	—
1883	12·04	11·27	+77	15·5	15·1	+4	308	—
1884	10·92	10·92	0	15·1	15·0	+1	0	—
1885	10·30	10·56	−26	14·5	14·7	−2	52	—
1886	9·63	10·25	−62	14·2	14·5	−3	186	—
1887	9·90	10·37	−47	14·4	14·5	−1	47	—
1888	10·51	10·55	− 4	14·4	14·7	−3	12	—
1889	11·50	10·93	+57	15·0	15·0	0	0	—
1890	11·22	11·17	+ 5	15·5	15·2	+3	15	—
1891	11·52	11·17	+35	15·6	15·2	+4	140	—
1892	11·12	10·97	+15	15·4	15·2	+2	30	—
1893	10·50	10·85	−35	14·7	15·1	−4	140	—
1894	10·50	10·78	−28	15·0	15·2	−2	56	—
1895	10·61	10·81	−20	15·0	15·3	−3	60	—
1896	11·15	11·03	+12	15·7	15·6	+1	12	—
1897	11·27	—	—	16·0	—	—	—	—
1898	11·64	—	—	16·2	—	—	—	—

To obtain the measurement by comparing differences, the table of which the first lines are given was completed.

	X	DX	D²X	Y	DY	D²Y	DX.DY	D²X.D²Y
		Imports.			Marriage Rate.			
1845	330	−15	—	172	0	—	0	—ᐧ
1846	315	+ 6	+21	172	−14	−14	−84	−294
1847	321	−30	−36	158	+ 1	+15	−30	−540
1848	291		+91	159		+ 2		+182

	DX	D²X	DY	D²Y
Average	15·74	1	−·19	·04
Standard Deviation . .	57·13	80	5·3	6·78

Sum of DX . DY = 8902. Sum of D²X . D²Y = 12076.

Hence r from first differences is ·60 and from second differences ·45.

Example 7.—In the experiment described on pp. 304–6, 1,000 sums were formed each of the number of letters in 10 words. Write A for the sum of the letters in the first 5 words, B for the sum of the second 5, so that $x = A + B$. After each 10 a further 5 words were taken, for the sum of whose letters we write C ; then y was taken as $B + C$. We have thus 1,000 pairs, for which the correlation coefficient should be $\frac{1}{2}$, with standard deviation ·024.

Actually the correlation coefficient was ·553, more than twice the standard deviation from the fraction expected. A possible explanation of this is in the want of complete independence discussed on p. 306. The coefficients for four separate groups of 250 (for which the standard deviation is ·047) were ·56, ·50, ·58, ·59.

The regression is nearly rectilinear in the central region from $x = 40$ to $x = 61$; outside these numbers there are less than 20 cases to one value of x, and the standard deviation of the average of an array is greater than 2, so that a comparison is not worth while. The standard deviations of the values of $\bar{y}$, included are from 1·3 to 2·0.

Value of x.	Average value of corresponding y's.	Interquartile range of the y's.	$\bar{y}_x$ deduced from equation.
40	48·3	10½	45·3
41	44·8	10	45·8
42	49·1	11	46·4
43	47·4	11	46·9
44	47·1	9	47·5
45	46·4	9½	48·0
46	48·9	12	48·5
47	46·4	10	49·1
48	50·2	11	49·6
49	51·1	12	50·2
50	51·5	11½	50·7
51	49·6	7	51·3
52	53·6	13	51·8
53	53·6	16	52·3
54	51·1	10	52·9
55	51·9	6	53·4
56	53·4	14	54·0
57	52·7	12	54·5
58	52·7	11	55·1
59	60·2	8	55·6
60	56	11	56·1
61	57·2	11	56·7

The interquartile range as calculated from theory ($\cdot$67 of $2\sigma\sqrt{1 - r^2}$), formulæ (26) and (107) is 10·5, to which the observed ranges approximate, their average being 10·75. The range appears to be independent of the value of x, as was to be expected from the theory (formula (107)).

The correspondence of these numbers is evident from the diagram, where the equation of the line of regression is

$$\frac{y - 51\cdot50}{9\cdot24} = \cdot553\frac{x - 51\cdot46}{9\cdot43}.$$

Number of letters experiment.

Averages of arrays and line of regression.

The straight line has for its equation

$$\frac{y-51.504}{9.244} = .853 \; \frac{x-51.455}{9.430}$$

The broken line shows the averages of the arrays; it is dotted below x=40 and above x=61 where there are less than 20 cases

The xx show the quartiles in each array.

Scale of x

We obtain a better numerical view of the regression if we group the numbers in wider grades, in which the error of sampling is diminished, thus :—

Grade of y.	Number of cases.	Average of y.	Average of corresponding x's.	x from equation.
30–39	85	36·3	43·8	42·9
40–49	348	44·7	47·7	47·6
50–59	360	54·3	52·7	52·0
60–69	173	63·0	57·7	58·0
70–79	30	72·7	64·1	63·4

Here the regression of x on y is taken ; in the diagram the regression is that of y on x.

There are two examples below 30 and two above 80.

There are various methods of comparing the distributions of observations with that given by the normal correlation surface, of which the simplest is as follows.

Take $r = \frac{1}{2}$ as given *a priori* and $\sigma = 9\cdot32$, the mean of the standard deviations of x and y.

The equation of the surface is

$$z = \frac{1}{2\pi\sigma^2\sqrt{1-r^2}} e^{-\frac{1}{2(1-r^2)\sigma^2}(x^2+y^2-xy)}$$

Write

$$x = \frac{X-Y}{\sqrt{2}}, \quad y = \frac{X+Y}{\sqrt{2}}.$$

The equation becomes $z = \dfrac{1}{\pi\sigma^2\sqrt{3}} \cdot e^{-\frac{X^2}{8\sigma^2}} \cdot e^{-\frac{Y^2}{\sigma^2}}$, and represents the surface referred to its principal axes, inclined at 45° to the original axes.

The volume standing on the area bounded by $X = X_1$, $X = X_2$, $Y = Y_1$, $Y = Y_2$ is

$$\frac{1}{\sqrt{2\pi} \cdot \sigma\sqrt{\frac{3}{2}}} \cdot \int_{X_1}^{X_2} e^{-\frac{X^2}{2(\sigma\sqrt{\frac{3}{2}})^2}} dX \cdot \frac{1}{\sqrt{2\pi} \cdot \sigma\sqrt{\frac{1}{2}}} \int_{Y_1}^{Y_2} e^{-\frac{Y^2}{2\left(\frac{\sigma}{\sqrt{2}}\right)^2}} dY,$$

and can be obtained at once from the table on p. 271.

.0003   P₂   .0034   2σ   K₂   .0112   L₂   .0106   Axis of x

N3   .0047   P₁ $\frac{3\pi}{2}$ .0215   L₁   .0282   M3

N2   .0296   σ   K₁   .0539   M2

N1 $\frac{\sigma}{2}$ .0744   M1

-2σ   -$\frac{3}{2}$σ   -σ   -$\frac{\sigma}{2}$ O $\frac{\sigma}{2}$ σ $\frac{3}{2}$σ   2σ   Axis of x

-$\frac{\sigma}{2}$

-σ

-$\frac{3\sigma}{2}$ Axis of y   Axis of Y

-2σ
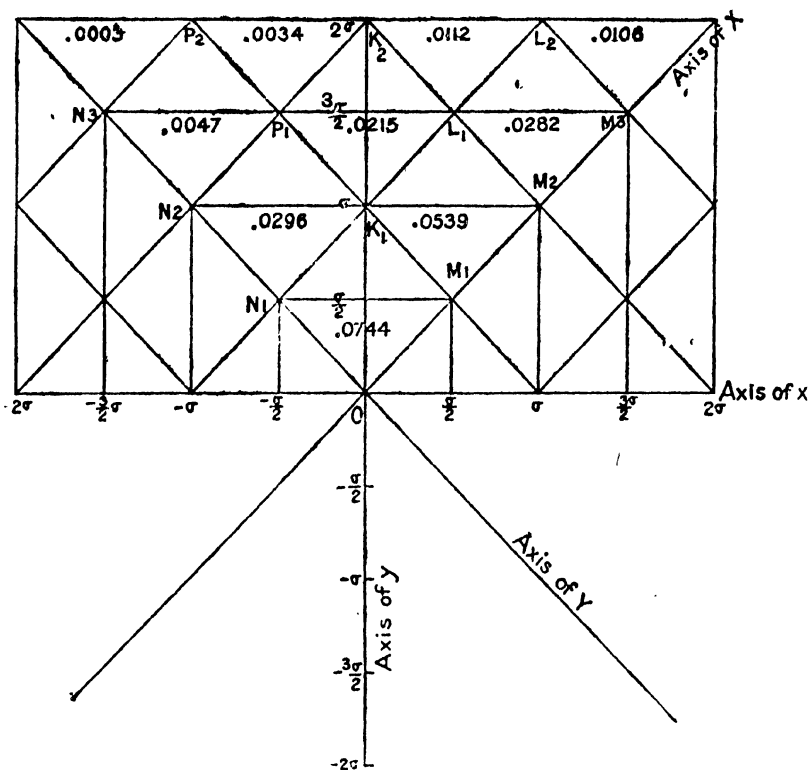
Mark distances on the axis of X as in the figure, OM_1, M_1M_2 . . . each equal to $\sigma/\sqrt{2}$. Write σ_1, σ_2 for the standard deviations of X and Y. $\sigma_1 = \sigma\sqrt{\frac{3}{2}}$, $\sigma_2 = \sigma/\sqrt{2}$.

Then $OM_1 = M_1M_2 = \ldots = \sigma_1/\sqrt{3} = \cdot 577\sigma_1$.

The proportional volumes of the solid bounded by vertical planes perpendicular to the axis of X are then F (·577) across OM_1 F (1·155) − F (·577) across M_1M_2, etc., which can be found from the table on p. 271, as ·2180, ·1580, etc.

Now mark distances ON_1, N_1N_2 . . . on the axis of Y each equal to $\sigma/\sqrt{2}$, that is to σ_2.

The proportional volumes bounded by vertical planes perpendicular to the axis of Y are F(1) across ON_1, F(2) − F(1) across ON_2, etc., viz. ·3413, ·1359, etc.

Since in the equation the integrals of X and Y are inde-

pendent, all sections perpendicular to the axis of X are cut in the same proportions by the planes through N_1, N_2, etc.

Hence we have the following table which shows the proportions of the volume of the normal surface standing on the squares $ON_1K_1M_1$, $M_1K_1L_1M_2$, $M_2L_1L_2M_3$ ... in the first line, on $N_1N_2P_1K_1$, $K_1P_1K_2L_1$ ... in the second line, etc.

DISTRIBUTION ON SQUARES OF NORMAL FREQUENCY SURFACE.

X/σ_1		·577	1·155	1·732	2·31	2·89	3·46
$F(X/\sigma_1)$ (differences) .		·2180	·1580	·0824	·0311	·0085	·0017

Y/σ_2	$F(Y/\sigma_2)$ Differences.	Products of Differences.					
1	·3413	·0744	·0539	·0282	·0106	·0028	·0006
2	·1359	·0296	·0215	·0112	·0042	·0012	·0002
3	·0214	·0047	·0034	·0017	·0007	·0002	·0000
4	·0014	·0003	·0002	·0001	·0001	·0000	·0000

The distribution is the same in each of the four quadrants formed by the axes OX and OY.

The decimals in this table × 1000 give the theoretic distribution of the 1,000 pairs of numbers if we neglect the skewness.

The observations were marked in on squared paper and the number occurring in each of the X,Y squares was counted.

The results are shown in the table on p. 394. The first line in each row repeats the theoretic numbers first given, the third gives the observations.

The agreement is close within the three squares to right and left, and two squares above and below the centre, that is within ± 1·7σ_1 and ± 2σ_2. The probability of so much divergence in a random selection is approximately ⅔ (p. 433).

To the left of these squares there is a falling-off in the observations (31 observations against 41 expected) and to the right an excess (54 observations against 41 expected). There is, however, a slight heaping up in the 12 squares to the left of the centre and a corresponding deficit to the right. These are exactly the phenomena we should expect from the skewness of the original curve (p. 304). The effect of the skewness is worked out in the note at the end of this chapter, and the results of the corrections are given in the second line of each row in the following table. The improvement is marked.

For example the expectation in the last three columns to the left is now 33½ (31 observations) and in the last three columns to the right is about 49 (54 observations).

1,000 *Pairs of Totals of Letters.*

Distribution of observations compared with normal and with skew frequency.

The central horizontal and vertical lines are not the axes of co-ordinates, but are the axes of symmetry, which are inclined at 45°.

First lines. Normal distribution . . (thus 29·6)
Second lines. Second approximation . (thus *28·7*)
Third lines. Observations . . . . (thus 35)

o	o	·1	·1	·2	·3	·3	·2	·1	·1	o	o
0	*0*	*?*	*?*	*0*	*·2*	*·4*	*·4*	*?*	*?*	*0*	*0*
0	**0**	**0**	**0**	**0**	**1**	**0**	**1**	**0**	**0**	**0**	**0**
o	·2	·7	1·7	3·4	4·7	4·7	3·4	1·7	·7	·2	o
?	*?*	*0*	*·2*	*1·9*	*4·3*	*5·1*	*4·9*	*3·2*	*1·7*	*?*	*0*
0	**0**	**0**	**1**	**1**	**1**	**5**	**6**	**4**	**0**	**0**	**0**
·2	1·2	4·2	11·2	21·5	29·6	29·6	21·5	11·2	4·2	1·2	·2
?	*·2*	*2·8*	*10·8*	*22·3*	*30·5*	*28·7*	*20·7*	*11·6*	*5·6*	*2·2*	*?*
0	**0**	**0**	**11**	**20**	**33**	**35**	**20**	**11**	**5**	**0**	**0**
·6	2·8	10·6	28·2	53·9	74·4	74·4	53·9	28·2	10·6	2·8	·6
0	*2·7*	*11·1*	*33·3*	*63·0*	*79·2*	*69·6*	*44·8*	*23·1*	*10·1*	*2·9*	*1·6*
0	**2**	**12**	**34**	**47**	**77**	**72**	**61**	**24**	**8**	**5**	**1**
·6	2·8	10·6	28·2	53·9	74·4	74·4	53·9	28·2	10·6	2·8	·6
0	*2·7*	*11·1*	*33·3*	*63·0*	*79·2*	*69·6*	*44·8*	*23·1*	*10·1*	*2·9*	*1·6*
0	**1**	**9**	**36**	**75**	**76**	**64**	**46**	**26**	**15**	**5**	**4**
·2	1·2	4·2	11·2	21·5	29·6	29·6	21·5	11·2	4·2	1·2	·2
?	*·2*	*2·8*	*10·8*	*22·3*	*30·5*	*28·7*	*20·7*	*11·6*	*5·6*	*2·2*	*?*
0	**1**	**1**	**8**	**21**	**32**	**26**	**18**	**9**	**8**	**0**	**2**
o	·2	·7	1·7	3·4	4·7	4·7	3·4	1·7	·7	·2	o
0	*?*	*0*	*·2*	*1·9*	*4·3*	*5·1*	*4·9*	*3·2*	*1·7*	*?*	—
0	**0**	**0**	**0**	**3**	**2**	**2**	**0**	**3**	**1**	**0**	**0**
o	o	·1	·1	·2	·3	·3	·2	·1	·1	o	o
0	*0*	*?*	*?*	*0*	*·2*	*·4*	*·4*	*?*	*?*	*0*	*0*
0	**0**	**0**	**0**	**0**	**0**	**2**	**1**	**0**	**0**	**0**	**0**

The probability of the divergence from expectation as a whole has been tested (see p. 433), and is approximately ⅖; that is, in only two such experiments out of five should we expect so close an agreement.* On the other hand, it is highly improbable that we should get so great divergence on the left and right if the distribution had been normal (and symmetrical) ;

---

* The thick lines in the table are only to mark the regions to which the test of p. 431 is applied.

the second approximation is necessary for the completion of the theory.

A simple method of testing the agreement of the distribution of observations with that given by the *normal* surface may be obtained by studying the distribution of the *x*-arrays, instead of transferring as on p. 391 to axes of symmetry.

$$\iint \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} e^{-\frac{1}{2(1-r^2)}\left(\frac{x^2}{\sigma_1^2}+\frac{y^2}{\sigma_2^2}-\frac{2rxy}{\sigma_1\sigma_2}\right)} dx\,dy$$

$$= \int \frac{1}{\sigma'\sqrt{2\pi}} e^{-\frac{y'^2}{2\sigma'^2}} dy' \times \int \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma_1^2}} dx,$$

where $\sigma' = \sigma_2\sqrt{(1-r^2)}$, $y' = y - \frac{rx\sigma_2}{\sigma_1}$, *i.e.*, y' is measured parallel to the axis of y from the line of regression, as in formula (107).
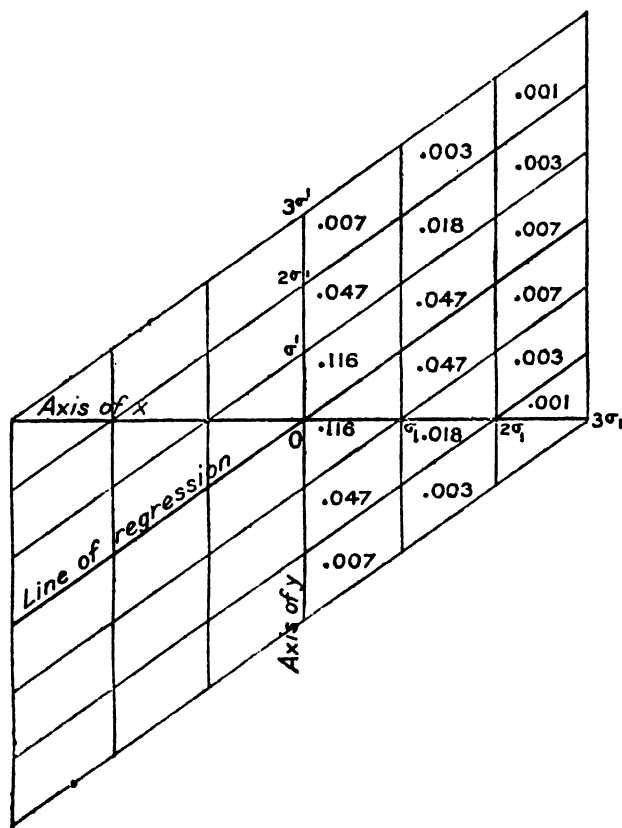
	.F(z)		
	0-1	1-2	2-3
	·341	·136	·022

F(z)	Products.			
0-1	·341	·116	·047	·007
1-2	·136	·047	·018	·003
2-3	·021	·007	·003	·001

The division when the standard deviations are taken as units is shown in the table and diagram.

The results of the words experiment tabulated on ·this basis are :—

	-3σ	-2σ	-σ	0	σ	2σ	3σ	
3σ′ calc	0	0	0	0	0	0	0	0
obs	0	0	0	0	0	0	0	0
2σ′ calc	0	1	3	7	7	3	1	0
obs	0	0	0	4	8	2	0	0
σ′ calc	0	3	18	47	47	18	3	0
obs	0	1	17	63	56	18	4	1
0 calc	1	7	47	116	116	47	7	1
obs	0	1	57	131	114	42	6	0
-σ′ calc	1	7	47	116	116	47	7	1
obs	0	7	49	98	103	48	9	1
-2σ′ calc	0	3	18	47	47	18	3	0
obs	1	2	13	47	50	16	0	0
-3σ′ calc	0	1	3	7	7	3	1	0
obs	0	2	3	10	7	3	0	0
calc	0	0	0	0	0	0	0	0
obs	0	0	1	0	3	1	0	1

{ Line of regression.

The vertical columns show the x-arrays, and are comparable with the more detailed setting out on p. 389. In each compartment the calculated values are written above the number of observations.

The effect of correcting for skewness would be to improve the correspondence in the same directions as in the former tabulation.

NOTE ON THE SECOND APPROXIMATION TO THE CORRELATION SURFACE

When terms of the order $\dfrac{1}{\sqrt{n}}$ are retained in the general law of great numbers, a term involving the mean cube of error appears in the equation

(p. 298).   Similarly Prof. Edgeworth shows [*] that the equation of the correlation surface should under similar conditions be written

$$z = z_0 - \tfrac{1}{6}\left(k_{30}\frac{\partial^3}{\partial x^3}z_0 + 3k_{21}\frac{\partial^3}{\partial x^2\partial y}z_0 + 3k_{12}\frac{\partial^3}{\partial y\partial x^2}z_0 + k_{03}\frac{\partial^3}{\partial y^3}z_0\right) \quad . \quad . \quad . \quad (113)$$

where

$$z_0 = \frac{1}{2\pi\sqrt{1-r^2}}e^{-\frac{1}{2(1-r^2)}\left(x^2+y^2-2rxy\right)},$$

x and y are the differences between the observations and their averages divided by their standard deviations,

$$k_{30} = \text{mean } x^3, \; k_{21} = \text{mean } x^2 y, \; k_{12} = \text{mean } xy^2, \; k_{03} = \text{mean } y^3.$$

In the example on pp. 388 *seq.* we have $x\sigma = A + B$, $y\sigma = B + C$, where $\sigma = 9.42$ approx.

Mean $xy\sigma^2 = $ mean $B^2 = \tfrac{1}{2}\sigma^2$, and $r = \tfrac{1}{2}$. (In the experiment r was found to be ·55.)

Mean $x^3\sigma^3 = k_{30} = k_{03} = \kappa$ (as on p. 251) $= \cdot409$.

Mean $x^2y\sigma^3 = k_{21} = $ mean $B^3 = \tfrac{1}{2}$ mean $(A + B)^3 = \tfrac{1}{2}\kappa = k_{12}$.

When the differentiations are performed with these values, we obtain

$$z = \frac{1}{\pi\sqrt{3}}e^{-\frac{2}{3}(x^2+y^2-xy)}\left\{1 - \cdot01\,(x+y)(18+11xy-8x^2-8y^2)\right\} = z_0(1-w) \text{ say.}$$

The expression z_0w is not readily integrable, and the simpler method of procedure is to integrate z_0 over suitable areas, and to correct the results approximately.   An application of the method that leads to Simpson's rule shows that if

$$z_{0,\,h}, \; z_{0,\,-h}, \; z_{h,\,0}, \; z_{-h,\,0},$$

are the ordinates of the four corners of a surface standing on a rectangular base, whose diagonals are $2h$, $2h$, and z_{00} is the ordinate at the centre of the rectangle, then the mean ordinate is

$$\tfrac{1}{6}\,(2z_{00} + z_{0k} + z_{0-k} + z_{h0} + z_{-h0})$$

Hence if z_0w is calculated for the four corners and centre of each of the volumes tabulated on p. 393, we should reduce each of the volumes by a quantity z_0w', where w' is the average of twice its central value and the four angular values.   This has been done throughout, and the values obtained added to or subtracted from the numbers given by the normal curve to obtain the corrected values on p. 394.

Notice that when the surface is referred to its principal axes by writing $x + y = \sqrt{2}\,X$, $- x + y = \sqrt{2}\,Y$, w becomes symmetrical in Y, but not in X.

---

[*] Law of Error (*Camb. Phil. Trans.*, Vol. XX, 1905, Part II, § 6), and *Statistical Journal*, 1917, pp. 268 *seq.* The standard deviation is used as unit in the text instead of the modulus $\sqrt{2}\cdot\sigma$ used by Edgeworth.

CHAPTER VIII.

PARTIAL AND MULTIPLE CORRELATION.

THE investigation of Chapter VII has shown how the variations in one quantity are related to the variations in another by which it is influenced. It frequently happens, however, that the movements of a variable are related to the movements of a number of others. The frequency distribution can then no longer be represented by a surface in three dimensions, but an analogous function is obtained of which the form already given is a simple case.

The regression equation is no longer that of a line or curve, but an expression connecting one selected variable ' with a number of others; we can then isolate the effect of any one of the remaining variables (by a method involving and similar to that of partial differentiation), and so obtain the relation between any pair of variables, abstraction being made of the remainder. This is the very important method of partial correlation.

In the sequel the case of three variables is handled in detail, and the more general solution is summarized.

Let there be three variables, which measured from their means are x, y, z, and let them be correlated each with each.

Suppose that they are so connected that $z = ax + by + c$ is an ideal plane giving the mean value of z corresponding to a pair of values x, y.

Required to find a, b, c so that the observed deviations of observed values of z from the values given by this equation have the least improbability.

Let z_s be the average of k_s observations, each of which has for its x, y members x_s (to $x_s + \delta x$) and y_s (to $y_s + \delta y$).

Write η_s for $z_s - (ax_s + by_s + c)$, *i.e.* the deviation of the mean of the observations in the sth group from its ideal value.

Let σ_z, σ_y, σ_s be the standard deviations of the frequency groups of x, y, z. Then, if in the long run the standard deviation of a group in z is independent of the values of x and y, the standard deviation of η_s is $\dfrac{\sigma}{\sqrt{k_s}}$ (formula (38)), where σ is constant standard deviation for an array, and the probability of the occurrence of η_s (to $\eta_s + \delta\eta$) is $Ke^{-\frac{k_s\eta_s{}^2}{2\sigma^2}} \cdot \delta\eta$.

Let there be n pairs of values such as x_s, y_s and N observations in all, so that $N = k_1 + \ldots + k_s + \ldots + k_n$.

The probability of the concurrence of $\eta_1 \ldots \eta_s \ldots \eta_n$ is $Ce^{-\frac{1}{2\sigma^2} \cdot \phi}$ where $\phi = k_1\eta_1{}^2 + \ldots + k_s\eta_s{}^2 + \ldots + k_n\eta_n{}^2$, and C is constant.

The probability is greatest when $\phi = \overset{n}{\underset{1}{S}} k_s[\bar{z}_s - (ax_s + by_s + c)]^2$ is least, and a, b, and c must be chosen to give this result.

$\phi = S(k_s\bar{z}_s{}^2) + a^2 S(k_s x_s{}^2) + b^2 S(k_s y_s{}^2) + c^2 Sk_s - 2a S(k_s x_s\bar{z}_s)$
$\quad - 2b S(k_s y_s\bar{z}_s) - 2c S(k_s\bar{z}_s) + 2ab S(k_s x_s y_s) + 2ac Sk_s x_s + 2bc Sk_s y_s.$

Here $Sx_s = 0 = Sy_s$. $Sk_s\bar{z}_s = $ sum of all values of $z = 0$.

$Sk_s x_s{}^2 = N\sigma_x{}^2$, x_s being repeated k_s times in the whole group, and $Sk_s y_s{}^2 = N\sigma_y{}^2$.

$Sk_s x_s\bar{z}_s = Sxz$, since $k_s\bar{z}_s = $ sum of values of z in the sth group, and $Sk_s y_s\bar{z}_s = Syz$. Also $Sk_s x_s y_s = Sxy$.

$\therefore \phi = S(k_s\bar{z}_s{}^2) + Na^2\sigma_x{}^2 + Nb^2\sigma_y{}^2 - 2aSxz - 2bSyz + 2abSxy + Nc^2.$

Write $Sxy = Nr_{xy}\sigma_x\sigma_y$, $Sxz = Nr_{xz}\sigma_x\sigma_z$, $Syz = Nr_{yz}\sigma_y\sigma_z$.

Then ϕ is a minimum, when $\dfrac{\partial\phi}{\partial a}$, $\dfrac{\partial\phi}{\partial b}$, $\dfrac{\partial\phi}{\partial c}$ are each zero,* $i.e.$, when

$$0 = \frac{\partial\phi}{\partial c} = 2Nc \quad \text{and} \quad \therefore c = 0,$$

$$0 = \frac{\partial\phi}{\partial a} = 2N(a\sigma_x{}^2 + b\sigma_x\sigma_y r_{xy} - \sigma_x\sigma_z r_{xz}) = 0,$$

and $\quad 0 = \dfrac{\partial\phi}{\partial b} = 2N(a\sigma_x\sigma_y r_{xy} + b\sigma_y{}^2 - \sigma_y\sigma_z r_{yz}) = 0.$

Hence $\quad\quad\quad\quad a\sigma_x + b\sigma_y r_{xy} = \sigma_z r_{xz}$
and $\quad\quad\quad\quad a\sigma_x r_{xy} + b\sigma_y = \sigma_z r_{yz}.$

$\therefore \quad \dfrac{a\sigma_x}{r_{xz} - r_{xy}r_{yz}} = \dfrac{b\sigma_y}{r_{yz} - r_{xz}r_{xy}} = \dfrac{\sigma_z}{1 - r_{xy}{}^2} \quad . \quad . \quad . \quad (114)$

---

* The values of a, b and c can be obtained without differentiation by expressing ϕ as the sum of squares.

The equation $z = ax + by + c$ becomes

$$\frac{z}{\sigma_z} = R_x \cdot \frac{x}{\sigma_x} + R_y \cdot \frac{y}{\sigma_y}, \quad \ldots \ldots \quad (115)$$

where $\qquad R_x = \dfrac{r_{xz} - r_{xy}r_{yz}}{1 - r_{xy}^2}, \quad R_y = \dfrac{r_{yz} - r_{xy}r_{xz}}{1 - r_{xy}^2}.$

R_x and R_y are called partial regression coefficients between z, x and z, y; for a given y, $z = R_x \dfrac{\sigma_z}{\sigma_x} x + \text{const.}$, and for a given x $z = R_y \dfrac{\sigma_z}{\sigma_y} \cdot y + \text{const.}$, formulæ which may be compared with $y = r \dfrac{\sigma_y}{\sigma_x} \cdot x$ given above (p. 362).

Similar equations are of course to be obtained when x or y are expressed in terms of y, z or x, z.

The partial *correlation* coefficient between x and z (y constant) is defined, by analogy with the case of only two variables, as the geometric mean of the partial regression coefficients found respectively when z is expressed in terms of x and y as in (115) and when x is expressed in terms of z and y; it is therefore

$$\frac{r_{xz} - r_{xy}\,r_{yz}}{\sqrt{1 - r_{xy}^2}\,\sqrt{1 - r_{yz}^2}}.$$

The foregoing analysis is based on Mr. Yule's paper (*Statistical Journal*, 1897, pp. 831 *seq.*) and book, and to him is due a great part of the work on this subject. The treatment here differs from his only by the important consideration that it is based on the prevalence of the law of error as discussed above (p. 298), and that it makes the assumption that the standard deviation of z is independent of the values of x and y, which is by no means universal ; while Mr. Yule does not need this assumption, but uses the method of least squares, a method which is not used (except very rarely) in this book, because of the difficulties that underlie the principles involved.

The equation between z and x and y is the same as Mr. Yule's, and also the same as obtained (see p. 405 below) from the theory of normal multiple correlation.

Example 1.—The Cost of Living Committee, 1918, collected a number of budgets of the weekly expenditure on food in working-class families (see p. 310) ; 390 of these, obtained from families of the skilled classes, were grouped accòrding to the

numbers of persons in the households above and below 14 ·years of age (see *Statistical Journal*, 1919, p. 360).

The notation and' quantities involved are as follows :—

	Expenditure on food.	Number over 14 years.	Number under 14 years.
Average	51*s*.	2·48	3·56
Difference from average .	*z* × 5*s*.	*x*	*y*
Standard deviation . .	$\sigma_z = 3\cdot03 \times 5s.$	$\sigma_x = \cdot836$	$\sigma_y = 1\cdot40$

$$r_{xy} = -\cdot0525, \quad r_{zx} = \cdot504, \quad r_{zy} = \cdot315.$$

The equation obtained is $z/\sigma_z = \cdot52x/\sigma_x + \cdot35y/\sigma_y$, which leads to the formula :—

Expenditure (shillings) $=14\cdot5 + 9\cdot4 \times$ number over 14 years $+ 3\cdot7 \times$ number of children under 14 years,

and to the following table :—

FAMILY EXPENDITURE ON FOOD.* (SHILLINGS.)

Number of persons over 14 years.	By Formula. Number of children.				Average of actual cases. Number of children.			
	2	3	4	5	2	3	4	5
2 . . .	40·7	44·4	48·1	51·8	40·5 (74)	45·2 (74)	47·1 (53)	52·9 (25)
3 . . .	50·1	53·8	57·5	61·2	54·8 (21)	51·2 (17)	58·2 (16)	64·9 (17)
4 . . .	59·5	63·2	66·9	70·6	58·0 (10)	60·2 (10)	78·1 (6)	— (0)

The numbers in brackets are the numbers of actual cases averaged.

The agreement between experience and formula is as close as can be expected, when the considerable standard deviation and the small numbers of cases are considered.

From this example it becomes clear that the method of partial correlation is closely akin to the ordinary way of analysis in cross-tabulation ; but the use of the formula brings the separate results into coherent relation. Here we have the result that on the average an additional adult (who generally increases family income) adds 9*s*. 5*d*. to the family

---

* For the working out of these figures and those on p. 310 I am indebted to Miss King and Miss Mackenzie at the School of Economics.

food 'expenditure, while an additional child adds only 3s. 8d. ; the greater number of children the lower is the standard of living, since a child's nourishment costs about two-thirds of that of an adult. (Here " adult " is used for a person over 14 years.)

Example 2.—The following data are obtained for the County of London from the Census of 1911.*

     $z + 3\cdot7$ is the number of rooms to a tenement.
     $x + 4\cdot15$,,   ,,   persons in a family.
     $y + \cdot86$,,   ,,   children under 10 in a family.

The averages for the county are 3·7, 4·15 and ·86 respectively.

$r_{xy} = \cdot57$, $r_{xz} = \cdot44$, $r_{yz} = -\cdot03$, $\sigma_z = 2\cdot59$, $\sigma_x = 2\cdot32$, $\sigma_y = 1\cdot24$, $R_x = \cdot676$, $R_y = -\cdot402$.

The figures relate to 1,023,951 families, sufficient for accuracy to three significant figures.

$$z = x \times \frac{\sigma_z}{\sigma_x} \times \cdot676 - y \times \frac{\sigma_z}{\sigma_y} \times \cdot402 = x \times \cdot754 - y \times \cdot840$$

or (rooms $- 3\cdot7$) $= \cdot754$ (persons $- 4\cdot15$) $- \cdot84$ (children $- \cdot86$).

The number of rooms for families of given size decreases rapidly as the number of children increases.

We may also write :—

    Number of rooms $= 1\cdot29 + \cdot75$ persons $- \cdot84$ children
                     $= 1\cdot29 + \cdot75$ persons over 10
                     $- \cdot09$ children under 10.

Example 3.—In the research on the social conditions in Reading described in *Livelihood and Poverty* the income, rent, and family constitution were tabulated for 586 families. Rent increases with income and with the number of earners, but for the same income and the same number of persons it may be that the more numerous the children the less can be afforded for rent.

Rent: $z + 6\cdot075$ shillings, where 6·075 shillings is the average.

Number of equivalent persons: $x + 3\cdot287$, when 3·287 is the average.

---

* I am indebted to Mr. J. W. Nixon for this calculation.

Income: $y + 31 \cdot 712$ shillings, where $31 \cdot 712$ shillings is the average.

The number of "equivalent" persons was obtained by classifying adults and children on a somewhat arbitrary scale according to the house-room they may be presumed to need; children under 5 years were counted as $\frac{1}{4}$, children from 5 to 14 as $\frac{1}{2}$, boys of 14 to 18 and girls from 14 to 16 as $\frac{3}{4}$, and older persons as 1.

The correlation between rent and number of rooms is close, so that rent may be taken as measuring house-room.

$\sigma_z = 1 \cdot 33$, $\sigma_x = 1 \cdot 22$, $\sigma_y = 13 \cdot 0$, $r_{xy} = \cdot 543$, $r_{xz} = \cdot 152$, $r_{yz} = \cdot 458$, $R_x = -\cdot 136$, $R_y = \cdot 532$.

Hence
$$\frac{z}{\sigma_z} = -\cdot 136 \frac{x}{\sigma_x} + \cdot 532 \frac{y}{\sigma_y},$$
or
$$z = -\cdot 148x + \cdot 0544y.$$

House-room then decreases perceptibly as the size of the family increases, for given incomes.

Each of the three examples shows that families with children tend to secure relatively less food and less house-room per head than families without children, and to some extent measures the loss.

We have still to consider the theoretic distribution in three dimensions of three variables, corresponding to the normal correlation surface for two variables. The following pages show the results and the analysis in simple cases. It will be observed that the same lines of proof are followed as in the case of two variables.

Multiple Correlation Surface.

The following analysis is only valid on the assumption that the elements have normal frequency.

Let X, Y, Z be three quantities which depend on other quantities U, V_1, V_2, V_3 in such a way that $X = U + V_1$, $Y = U + V_2$, $Z = U + V_3$.

Let U, V_1, V_2, V_3 be chosen at random from normal frequency groups whose averages are $\bar{u}$, $\bar{v}_1$ ... and standard deviations σ_u, σ_{v_1} ...

Let $\overline{X}$, $\overline{Y}$, $\overline{Z}$ be the averages and σ_x, σ_y, σ_z the standard deviations of X, Y, Z, and let $X = \overline{X} + x \ldots$ $U = \overline{U} + u \ldots$

Then in the long run $\overline{X} = \overline{U} + \overline{V}$ etc. and

$$\therefore \qquad x = u + v_1, \quad y = u + v_2, \quad z = u + v_3.$$

Suppose u, v_1, v_2, v_3 quite independent of each other, so that mean $uv_1 = 0 =$ mean v_1v_2 etc.

Write r_{xy} for the coefficient of correlation between x and y.

Then
$$\sigma_x^2 = \sigma_u^2 + \sigma^2 \ldots,$$

and
$$\sigma_x\sigma_y r_{xy} = \text{mean} \, (u + v_1)(u + v_2) = \sigma_u^2$$
$$= \sigma_y\sigma_z r_{yz} = \sigma_z\sigma_x r_{zx}.$$

The joint chance of selected values of x_1, y_1, z_1 arising from particular values u, v_1, v_2, v_3 is

$$\frac{1}{\sigma_u . \sqrt{2\pi}} e^{-\frac{1}{2}\frac{u^2}{\sigma_u^2}} \times \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{1}{2}\frac{v_1^2}{\sigma_{v_1}^2}} \times \quad \times \quad = P_u,$$

subject to the conditions $x_1 = u + v_1$, $y_1 = u + v_2$, $z_1 = u + v_3$.

Eliminate v_1, v_2, v_3.

The joint chance of x_1, y_1, z_1 arising from a particular value u is given by

$$- 2 \log (P_u . 4\pi^2\sigma_u\sigma_{v_1}\sigma_{v_2}\sigma_{v_3})$$
$$= \frac{u^2}{\sigma_u^2} + \frac{(u - x_1)^2}{\sigma_{v_1}^2} + \frac{(u - y_1)^2}{\sigma_{v_2}^2} + \frac{(u - z_1)^2}{\sigma_{v_3}^2} = a(u - b)^2 + c,$$

where
$$a = \frac{1}{\sigma_u^2} + \frac{1}{\sigma_{v_1}^2} + \frac{1}{\sigma_{v_2}^2} + \frac{1}{\sigma_{v_3}^2}$$

$$ab = \frac{x_1}{\sigma_{v_1}^2} + \frac{y_1}{\sigma_{v_2}^2} + \frac{z_1}{\sigma_{v_3}^2}$$

$$c = \frac{x_1^2}{\sigma_{v_1}^2} + \frac{y_1^2}{\sigma_{v_2}^2} + \frac{z_1^2}{\sigma_{v_3}^2} - ab^2.$$

The whole chance of the selected values $x_1 \, y_1 \, z_1$ arising from *any* value of u is $P = \int_{-\infty}^{\infty} P_u \, du$, $x_1 \, y_1 \, z_1$ being regarded as constant

$$= \frac{e^{-\frac{c}{2}}}{(2\pi)^{\frac{3}{2}}\sigma_u\sigma_{v_1}\sigma_{v_2}\sigma_{v_3}} \int_{-\infty}^{\infty} \frac{e^{-\frac{a}{2}(u - b)^2}}{\sqrt{2\pi}} \, du = \frac{1}{(2\pi)^{\frac{3}{2}}\sigma_u\sigma_v \, \sigma_{v_2}\sigma_{v_3}} . \frac{1}{\sqrt{a}} e^{-\frac{c}{2}}.$$

Write x, y, z for x_1, y_1, z_1.

$$ac = \left(\frac{1}{\sigma_u^2} + \frac{1}{\sigma_{v_1}^2} + \ldots\right)\left(\frac{x^2}{\sigma_{v_1}^2} + \ldots\right) - \left(\frac{x}{\sigma_{v_1}^2} + \ldots\right)^2.$$

Now

$$\frac{1}{\sigma_u^2} + \frac{1}{\sigma_{v_2}^2} + \frac{1}{\sigma_{v_3}^2} = \frac{1}{\sigma_u^2} + \frac{1}{\sigma_y^2 - \sigma_u^2} + \frac{1}{\sigma_z^2 - \sigma_u^2} = \frac{\sigma_y^2\sigma_z^2 - \sigma_u^4}{\sigma_u^2\sigma_{v_2}^2\sigma_{v_3}^2}$$

and

$$a = \frac{1}{\sigma_u^2} + \frac{1}{\sigma_x^2 - \sigma_u^2} + \frac{1}{\sigma_y^2 - \sigma_u^2} + \frac{1}{\sigma_z^2 - \sigma_u^2}$$

$$= \frac{\sigma_x^2\sigma_y^2\sigma_z^2 - \sigma_u^4(\sigma_x^2 + \sigma_y^2 + \sigma_z^2) + 2\sigma_u^6}{\sigma_u^2\sigma_{v_1}^2\sigma_{v_2}^2\sigma_{v_3}^2}.$$

$$\therefore c\{\sigma_x^2\sigma_y^2\sigma_z^2 - \sigma_u^4(\sigma_x^2 + \sigma_y^2 + \sigma_z^2) + 2\sigma_u^6\}$$
$$= x^2(\sigma_y^2\sigma_z^2 - \sigma_u^4) + + \quad - 2xy\sigma_u^2\sigma_{v_3}^2 - \quad -.$$

Write $R\sigma_x^2\sigma_y^2\sigma_z^2$ for $\sigma_x^2\sigma_y^2\sigma_z^2 - \sigma_u^4(\sigma_x^2 + \quad +) + 2\sigma_u^6$,

so that $$R = 1 + 2r_{xy}r_{yz}r_{zx} - r_{xy}^2 - r_{yz}^2 - r_{zx}^2.$$

The chance of the concurrence of x, y, z is then

$$P = \frac{1}{(2\pi)^{\frac{3}{2}}\sigma_x\sigma_y\sigma_z R^{\frac{1}{2}}} \cdot e^{-\frac{1}{2R}\left\{\frac{x^2}{\sigma_x^2}(1 - r_{yz}^2) + + \quad - \frac{2xy}{\sigma_x\sigma_y}(r_{xy} - r_{zx}r_{yz}) - \quad -\right\}}. \quad (116)$$

for $$\frac{r_{xy} - r_{xz}r_{yz}}{\sigma_x\sigma_y} = \frac{\sigma_x\sigma_y r_{xy}\sigma_z^2 - \sigma_x\sigma_z \cdot r_{xy} \cdot \sigma_y\sigma_z \cdot r_{yz}}{\sigma_x^2\sigma_y^2\sigma_z^2}$$

$$= \frac{\sigma_u^2\sigma_z^2 - \sigma_u^2 \cdot \sigma_u^2}{\sigma_x^2\sigma_y^2\sigma_z^2} = \frac{\sigma_u^2\sigma_{v_3}^2}{\sigma_x^2\sigma_y^2\sigma_z^2}.$$

In the special case where

$$\sigma_u = \sigma_{v_1} = \sigma_{v_2} = \sigma_{v_3}, \quad \sigma_x^2 = 2\sigma_u^2 = \sigma^2, \text{ say.}$$
$$r_{xy} = \tfrac{1}{2} = r_{yz} = r_{zx} \quad \text{and} \quad R = \tfrac{1}{2}.$$

The chance is then

$$\frac{1}{2\sigma^3\pi^{\frac{3}{2}}} e^{-\frac{1}{4\sigma^2}\{3(x^2 + y^2 + z^2) - 2(xy + yz + zx)\}}.$$

The most probable value of z for given values of x and y is obtained from $\frac{\partial P}{\partial z} = 0$, and is

$$\frac{z}{\sigma_z}(1 - r_{xy}^2) = \frac{x}{\sigma_x}(r_{xz} - r_{xy}r_{yz}) + \frac{y}{\sigma_y}(r_{yz} - r_{xy}r_{zx}),$$

as in formula (115).

In the special case this becomes $z = \tfrac{1}{3}(x + y)$.

It is shown by Elderton (following Pearson) that if x, y and z are the sum of any finite number of variables, such as the $u, v, \ldots$ above, all of normal frequency, and some common to a pair (x, y) (y, z) or (z, x) and others occurring in only one of these, then P is of the form $Ke^{-(ax^2 + by^2 + cz^2 + 2fyz + 2gzx + 2hxy)}$ where a, b, c, f, g, h are constants to be determined.

Take the aggregate of the chances to be unity.

Let A, B, C, F, G, H be the minors of the determinant $\Delta = \begin{vmatrix} a & h & g \\ h & b & f \\ g & f & c \end{vmatrix}$, so

that $A = bc - f^2$, $F = hg - af \ldots$, $BC - F^2 = a\Delta$, ...

Then $-\log \dfrac{P}{K} = a\left(x + \dfrac{h}{a}y + \dfrac{g}{a}z\right)^2 + \dfrac{C}{a}\left(y - \dfrac{F}{C}z\right)^2 + \dfrac{\Delta}{C}z^2$

$1 = \iiint P \, dx \, dy \, dz = (\sqrt{\pi})^3 \cdot K \dfrac{1}{\sqrt{a}} \cdot \sqrt{\dfrac{a}{C}} \sqrt{\dfrac{C}{\Delta}}$ and $K\pi^{\frac{3}{2}} = \Delta^{\frac{1}{2}}$

$\sigma_x^2 = \iiint Px^2 \, dx \, dy \, dz = K\pi \cdot \dfrac{1}{\sqrt{a}} \cdot \sqrt{\dfrac{a}{C}} \int z^2 e^{-\frac{\Delta}{C}z^2} \, dz = K\pi^{\frac{1}{2}}\dfrac{1}{\sqrt{C}} \cdot \frac{1}{2}\left(\dfrac{C}{\Delta}\right)^{\frac{3}{2}} = \dfrac{C}{2\Delta}$

Similarly $\sigma_y^2 = \dfrac{B}{2\Delta}, \quad \sigma_z^2 = \dfrac{A}{2\Delta}$

$\sigma_y\sigma_z r_{yz} = \iiint Pzy \, dx \, dy \, dz = K\sqrt{\pi}a^{-\frac{1}{2}} \iint zy e^{-\frac{C}{a}\left(y - \frac{F}{C}z\right)^2} e^{-\frac{\Delta}{C}z^2} \, dy \, dz$

$= K\sqrt{\pi}a^{-\frac{1}{2}} \iint z\left(y' + \dfrac{F}{C}z\right)e^{-\frac{C}{a}y'^2} e^{-\frac{\Delta}{C}z^2} \, dy' dz,$

where $y' = y - \dfrac{F}{C}z$ and the limits of the integration are $\pm \infty$

$= K\pi \cdot a^{-\frac{1}{2}} \sqrt{\dfrac{a}{C}} \cdot \dfrac{F}{C}\int z^2 e^{-\frac{\Delta}{C}z^2} \, dz = \frac{1}{2}\dfrac{F}{\Delta}.$

Similarly $\sigma_x\sigma_y r_{xy} = \frac{1}{2}\dfrac{H}{\Delta}$ and $\sigma_x\sigma_z r_{xz} = \frac{1}{2}\dfrac{G}{\Delta}.$

$\therefore a\Delta = BC - F^2 = 4\sigma_y^2\sigma_z^2(1 - r_{yz}^2)\Delta^2$

$f\Delta = GH - AF = 4\sigma_x^2\sigma_y\sigma_z(r_{xy}r_{xz} - r_{yz})\Delta^2$

$\Delta^2 = ABC + 2FGH - AF^2 - BG^2 - CH^2$

$= 8\Delta^3\sigma_x^2\sigma_y^2\sigma_z^2(1 + 2r_{xy}r_{yz}r_{xz} - r_{yz}^2 - r_{xz}^2 - r_{xy}^2).$

Write R for the quantity inside the bracket. $R = \begin{vmatrix} 1 & r_{xy} & r_{xz} \\ r_{xy} & 1 & r_{yz} \\ r_{xz} & r_{yz} & 1 \end{vmatrix}$

Then $\Delta = \dfrac{1}{8R\sigma_x^2\sigma_y^2\sigma_z^2}, \quad a = \dfrac{1 - r_{yz}^2}{2R\sigma_x^2}, \quad f = \dfrac{r_{xy}r_{xz} - r_{yz}}{2\sigma_y\sigma_z R}.$

Hence $P = \dfrac{1}{\sqrt{R}(2\pi)^{\frac{3}{2}}\sigma_x\sigma_y\sigma_z} \cdot e^{-\frac{1}{2R}\left\{\frac{x^2}{\sigma_x^2}(1 - r_{yz}^2) + \ + \ -\frac{2yz}{\sigma_y\sigma_z}(r_{yz} - r_{xy}r_{xz}) - \ - \right\}}$

as obtained in the special case above.

If instead of x, y, z there are n quantities $x_1 \, x_2 \ldots x_n$ a more general proof on the same lines (due to Prof. Karl Pearson) leads to

$$P = \dfrac{1}{\sqrt{R}(2\pi)^{\frac{n}{2}} \cdot \sigma_1\sigma_2 \ \ \sigma_n} e^{-\frac{1}{2R}\left\{\frac{x_1^2}{\sigma_1^2}R_{11} + \ + \ + 2\frac{x_1 r_2}{\sigma_1\sigma_2}R_{12} + \ + \right\}}$$

where $R = \begin{vmatrix} 1 & r_{12} & r_{13} \dots r_{1n} \\ r_{21} & 1 & r_{23} \dots r_{2n} \\ \cdot & \cdot & \cdot \cdot \cdot \\ r_{n1} & r_{n2} & r_{n3} \dots 1 \end{vmatrix}$, r_{st} being the same quantity as r_{ts}, and

R_{st} being the minor obtained by crossing out the sth column and t^{th} row, with its appropriate sign.

The following note indicates the course of the proof.

As before $P = Ke^{-\phi}$, where $\phi = a_{11}x_1^2 + \dots + 2a_{12}x_1x_2 + \dots$, $a_{11} \dots a_{12} \dots$ being constants.

Let $\Delta_n = \begin{vmatrix} a_{11} & a_{12} \dots a_{1n} \\ a_{21} & a_{22} \dots a_{2n} \\ \cdot & \cdot \cdot \cdot \cdot \\ \cdot & \cdot \cdot \cdot \cdot \\ a_{n1} & a_{n2} \dots a_{nn} \end{vmatrix}$, where $a_{st} = a_{ts}$, and let A_{st} be a minor

obtained by crossing out the sth column and t^{th} row.

$$A_{nn} = \Delta_{n-1}.$$

Then ϕ may be expanded so that x_1 occurs only in the first term, x_2 only in the first two terms, etc., and then

$$\phi = a_{11}\left(x_1 + \frac{a_{12}}{a_{11}}x_2 + \dots\right)^2 + \frac{a_{22}a_{11} - a_{12}^2}{a_{11}}\left(x_2 + \dots\right)^2 + \dots + \frac{\Delta_{n-1}}{\Delta_{n-2}}\left(x_{n-1} - \frac{A_{n,n-1}}{\Delta_{n-1}}x_n\right)^2$$
$$+ \frac{\Delta_n}{\Delta_{n-1}} x_n^2,$$

as may be shown by a rather troublesome induction.

$$1 = \int\int \dots P \cdot dx_1 dx_2 \dots dx_n = K\pi^{\frac{n}{2}} \cdot \left(\frac{1}{a_{11}}\right)^{\frac{1}{2}}\left(\frac{\Delta_1}{\Delta_2}\right)^{\frac{1}{2}} \dots \left(\frac{\Delta_{n-1}}{\Delta_n}\right)^{\frac{1}{2}} = \frac{K\pi^{\frac{n}{2}}}{(\Delta_n)^{\frac{1}{2}}}$$

$$\sigma_n^2 = \int\int \dots Px_n^2 dx_1 dx_2 \dots dx_n = 1 \cdot \frac{1}{2}\frac{\Delta_{n-1}}{\Delta_n} = \frac{A_{nn}}{2\Delta_n}$$

Similarly, by changing the order, $\sigma_t^2 = \frac{A_{tt}}{2\Delta_n}$

$$\sigma_n \cdot \sigma_{n-1} \cdot r_{n,\,n-1} = \int\int \dots Px_n x_{n-1} dx_1 \dots dx_n$$

$$= \frac{1}{\pi}\left(\frac{\Delta_n}{\Delta_{n-2}}\right)^{\frac{1}{2}}\int\int x_n x_{n-1} e^{-\frac{\Delta_{n-1}}{\Delta_{n-2}}\left(x_{n-1} - \frac{A_{n,\,n-1}}{\Delta_{n-1}}x_n\right)^2 - \frac{\Delta_n}{\Delta_{n-1}}x_n^2} dx_{n-1}dx_n = \frac{A_{n,\,n-1}}{2\Delta_n}.$$

Similarly, by changing the order,

$$\sigma_s\sigma_t r_{st} = \frac{A_{st}}{2\Delta_n}.$$

Substitute these values for r_{st} etc. in the determinant giving R, and we obtain

$$R = \frac{1}{(2\Delta_n)^n\sigma_1^2 \dots \sigma_n^2}\begin{vmatrix} A_{11} & A_{12} \dots A_{1n} \\ A_{21} & A_{22} \dots A_{2n} \\ \cdot & \cdot \cdot \cdot \cdot \\ A_{n1} & A_{n2} \dots A_{nn} \end{vmatrix} = \frac{\Delta_n^{n-1}}{(2\Delta_n)^n\sigma_1^2 \dots \sigma_n^2},$$

by a well-known theorem in determinants

$$\frac{1}{R} = (2\pi)^{\frac{n}{2}}\sigma_1 \dots \sigma_n \cdot \sqrt{R}.$$

$$R_{11} = \frac{1}{(2\Delta_n)^{n-1}\sigma_2{}^2\ldots\sigma_n{}^2} \begin{vmatrix} A_{22} & \ldots & A_{2n} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ A_{n2} & \ldots & A_{nn} \end{vmatrix} = \frac{a_{11}\Delta_n{}^{n-2}}{(2\Delta_n)^{n-1}\sigma_2{}^2\ldots\sigma_n{}^2}, \quad \text{by another}$$

well-known theorem.

$$\therefore \qquad\qquad\qquad a_{11} = \frac{R_{11}}{2R\sigma_1{}^2}.$$

Similarly $a_{12} = \dfrac{R_{12}}{2R\sigma_1\sigma_2}$, and by changing the order $a_{st} = \dfrac{R_{st}}{2R\sigma_s\sigma_t}$, and hence we obtain the formula as given at the bottom of p. 406.

The most probable value of x_1 for given values of $x_2, x_3 \ldots x_n$ is then given by

$$\frac{x_1}{\sigma_1} \cdot R_{11} = -\frac{x_2}{\sigma_2}R_{12} - \frac{x_3}{\sigma_3}R_{13}\ldots - \frac{x_n}{\sigma_n} \cdot R_{1n}.$$

CHAPTER IX.

PRECISION OF MEASUREMENTS OF AVERAGES, MOMENTS AND CORRELATION.*

Inverse Probability.

IN the previous chapters the problems of the errors that arise in the process of sampling have been chiefly discussed from the point of view of the universe, not of the sample ; that is, the question has been how far will a sample represent a given universe ? The practical question is, however, the converse : what can we infer about a universe from a given sample ? This involves the difficult and elusive theory of inverse probability, for it may be put in the form, which of the various universes from which the sample may *a priori* have been drawn may be expected to have yielded that sample ?

To make the argument clear it seems expedient to make a short digression on the theory of inverse probability. The following examples illustrate the problem and its solution.

A sovereign and two shillings were in a purse. One coin is lost. One of the remaining two is taken out and is found to be a shilling. What is the chance that the sovereign was lost ?

The *a priori* chance that the sovereign was lost is $p'_1 = \frac{1}{3}$, and that a shilling was lost $p'_2 = \frac{2}{3}$, if we assume that the loss of any one coin was as likely as any other.

If the sovereign was lost, the chance of drawing a shilling was $p_1 = 1$, since there is no other to draw.

---

* See Edgeworth in *Statistical Journal*, 1908, pp. 381 *seq.* ; Yule, *Introduction to the Theory of Statistics*, last chapter ; *Transactions of the Royal Society,* Pearson and Filon, Vol. 191 (A. 220), and Sheppard, Vol. 192 (A. 229), 1898 ; *Biometrika*, Vol. II, Part III, p. 280.

If 'a shilling was lost, the chance of drawing a shilling was $p_2 = \frac{1}{2}$.

The *a priori* chance that it should be a sovereign that is lost and a shilling that is drawn, is $p_1'p_1 = \frac{1}{3}$.

The *a priori* chance that it should be a shilling that is lost and a shilling that is drawn is $p_2'p_2 = \frac{2}{3} \times \frac{1}{2} = \frac{1}{3}$.

By hypothesis one of these equally probable double events has happened, and there is nothing in the data to show which.

It is therefore just as likely that the third coin is a sovereign as a shilling.

We may generalise this proposition in the following 'way. Of various possible events whose chances of occurrence are respectively $p_1', p_2' \ldots p_t' \ldots$ one is known to have taken place. A further result is found, whose probability, if the first, second $\ldots t^{\text{th}} \ldots$ event had happened, would have been $p_1, p_2 \ldots p_t \ldots$ respectively.

The *a priori* chance that the tth event happened and produced the result is $p_t' \times p_t$.

A priori the chances that the events of the first series would happen and produce the result are in the ratios

$$p_1'p_1 : p_2'p_2 : \ldots : p_t'p_t \ldots = P_1 : P_2 : \ldots$$

But we know that one or other of these did happen, and this additional knowledge does not affect the *relative* magnitudes $P_1, P_2 \ldots$, but raises their total in such a ratio, K, that it equals 1, which represents certainty on the scale of algebraic chance. Hence $K.SP_t = 1$, and the chance that it was the tth event in the first series is

$$KP_t = \frac{p_t'p_t}{p_1'p_1 + \ldots + p_t'p_t + \ldots}$$

In a bag there are six similar balls which are known to be black or white. One is drawn and is found to be white. What can be inferred as to the original number of white balls in the bag ?

The answer depends on the hypothesis made as to the *a priori* chances of distribution between black and white.

If *a priori* each ball is equally likely to be black or white and p_t' is the chance that t were white, $p_t' = \frac{1}{2^6} \cdot {}_6C_t$.

$p_t = \dfrac{t}{6}$ whatever the hypothesis.

$$\therefore \quad P_0 : P_1 : \ldots : P_6 = \dfrac{1}{6 \times 2^6}\,(0 : 6 : 30 : 60 : 60 : 30 : 6).$$

$$SP_t = \tfrac{1}{2} \text{ and } K = 2.$$

The chances that there were 0, 1, 2, 3, 4, 5, 6 white balls are respectively 0, $\tfrac{1}{32}$, $\tfrac{5}{32}$, $\tfrac{10}{32}$, $\tfrac{10}{32}$, $\tfrac{5}{32}$, $\tfrac{1}{32}$.

But if the number of white in the bag had been determined by throwing a die and taking the number on the upper face, then

$$p_1' = p_2' = \ldots = p_6' = \dfrac{1}{6}; \quad P_t = \dfrac{1}{6}\cdot\dfrac{t}{6};$$

$$SP_t = \dfrac{21}{36}, \; K = \dfrac{12}{7}, \text{ and } KP_t = \dfrac{t}{21}.$$

More generally if there were n balls in the bag and the number of white had been determined by spinning a disc, marked on its circumference with the numbers 0, 1 . . . n equally spaced, on a vertical axis, and taking the number nearest a fixed point adjacent to its circumference when it came to rest, then

$$p_t' = \dfrac{1}{n+1}, \; p_t = \dfrac{t}{n}, \; SP_t = \dfrac{1}{2}, \; KP_t = \dfrac{2t}{n(n+1)}.$$

The aggregate chance that originally the number of white balls was t or less is

$$K(P_0 + P_1 + P_2 + \ldots + P_t) = \dfrac{t(t+1)}{n(n+1)} = f(t), \text{ say.}$$

If $f(t) = \tfrac{1}{2}$, it is as likely as not that there were as many as t white; and, if n is great, $t = \dfrac{n}{\sqrt{2}}$ satisfies this equation approximately.

Hence, when n is great, it is as likely as not that the proportion of white balls to all was as great as $\dfrac{1}{\sqrt{2}} = \cdot 7 \ldots$

The chance is approximately $\tfrac{1}{2}$ that the proportion was between $\tfrac{1}{2}$ and $\dfrac{\sqrt{3}}{2}$.

This example is very important, both as showing that the result depends on the hypothesis made as to the relative

a priori chances of the unknown events between which we have to choose, and as indicating that we can get a more comprehensive result by aggregating the chances than by taking them singly.

Precision of p, *the Proportion of a Particular Class in a Universe.*

We will first apply the principle of inverse probability to the determination by sample of p, where pN is the number of things having a certain characteristic in a universe containing N things, and n are selected at random and $p'n$ are found to have the characteristic.

The chance that $p'n$ should have been found from a given p is $_nC_{p'n}\, p^{p'n}\, q^{q'n}$ (p. 262) where $q=1-p$, $q'=1-p'$.

If all values of p *from* 0 *to* 1 *are* a priori *equally probable* then the chance that $p'n$ should be found from any value of p, from p' to x, is the sum of the chances from particular values $=\,_nC_{p'n}\int_{p'}^{x} x^{p'n}(1-x)^{q'n}\,dx$, and therefore by the theorem on p. 410 the chance that the original value of p was between p' and x is

$$P_x = \frac{_nC_{p'n} \cdot \int_{p'}^{x} x^{p'n}(1-x)^{q'n}\,dx}{_nC_{p'n} \cdot \int_{0}^{1} x^{p'n}(1-x)^{q'n}\,dx},$$

which can be reduced as follows to the form of the normal curve of error if $\dfrac{1}{\sqrt{n}}$ is neglected.

Write $x = p' + z\sigma$, where $\sigma^2 n = p'q'$, and $1 - x = q' - z\sigma$. σ is of order $\dfrac{1}{\sqrt{n}}$.

Then

$$P_x = \frac{\int_{0}^{z}\left(1 + \frac{z\sigma}{p'}\right)^{p'n}\left(1 - \frac{z\sigma}{q'}\right)^{q'n} dz}{\int_{-\infty}^{\infty}\left(1 + \frac{z\sigma}{p'}\right)^{p'n}\left(1 - \frac{z\sigma}{q'}\right)^{q'n} dz} = \frac{\int_{0}^{z} f(z)\,dz}{\int_{-\infty}^{\infty} f(z)\,dz}, \text{ say,}$$

since if $x = 1$, $z = \sqrt{\dfrac{nq'}{p'}}$, and if $x = 0$, $z = -\sqrt{\dfrac{np'}{q'}}$, which tend to $\pm \infty$ when n is great.

Then

$$\log f(z) = p'n \log\left(1 + \frac{z\sigma}{p'}\right) + q'n \log\left(1 - \frac{z\sigma}{q'}\right)$$

$$= z \times 0 - \frac{z^2\sigma^2 n}{2}\left(\frac{1}{p'} + \frac{1}{q'}\right) + \text{terms involving } \sigma^3 n,$$

i.e. terms of order $\dfrac{1}{\sqrt{n}}$

$$= -\tfrac{1}{2}z^2, \text{ when } \frac{1}{\sqrt{n}} \text{ is neglected, since } p' + q' = 1.$$

$$\therefore \mathrm{P}_x = \frac{\displaystyle\int_0^z e^{-\frac{1}{2}z^2}dz}{\displaystyle\int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2}dz} = \int_0^z \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}dz \quad \ldots \ldots \quad (117)$$

Hence the chance that the observations arose from a universe in which the proportion was between p' and $p' + p_1$ is (writing $\dfrac{p_1}{\sigma}$ for z) $\displaystyle\int_0^{p_1} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{u^2}{2\sigma^2}} du$, where $\sigma = \sqrt{\dfrac{p'(1-p')}{n}}$.

The above analysis is based on that given in Todhunter's *History of the Theory of Probability*, pp. 554 *seq.*

This is the converse of the theorem that the chance of obtaining a value from p to $p + p_1$ in a sample from a *known* universe is

$$\int_0^{p_1} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{u^2}{2\sigma^2}} du, \text{ where } \sigma = \sqrt{\frac{p(1-p)}{n}}.$$

The difficulty in the above analysis lies in the assumption that all values of p from 0 to 1 are *a priori* equally probable. The hypothesis can be elucidated as follows.

Let $n = 100$ and $p' = \cdot 1$.

If the observations came from a universe where $p = \cdot 07$, then $\sigma^2 = \dfrac{\cdot 07 \times \cdot 93}{100}$, $\sigma = \cdot 0255$. $\dfrac{p' - p}{\sigma} = 1 \cdot 18 = z$. The aggregate chance from $p = \cdot 06$ to $p = \cdot 08$ is approximately the chance for $p = \cdot 07$, exactly given by the ordinate $\dfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$, multiplied by an abscissa of $\cdot 02$ taken as a multiple of σ, viz. $\cdot 02 \div \cdot 0255 = \cdot 78$, and equals $\cdot 78$ of $\dfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(1\cdot 18)^2} = \cdot 157$.

A series of such calculations leads to the following table.

Value of p.	Approximate chance of obtaining $p' = \cdot 1$.
·00—·02	·000
·02—·04	·002
·04—·06	·029
·06—·08	·157
·08—·10	·262
·10—·12	·242
·12—·14	·159
·14—·16	·084
·16—·18	·039
·18—·20	·014
·20—·22	·005
·22—·24	·001
·24—·26	·000
	·994

From the observations p would be given as ·1 with standard deviation $\sqrt{\dfrac{\cdot 1 \times \cdot 9}{100}} = \cdot 03$, and a considerable positive skewness.

The values of p differing from ·1 by more than twice the standard deviation give negligible probabilities whether we suppose them *a priori* equally probable throughout the scale or not.

This example then illustrates a theorem that we may give as obvious : that, except in the neighbourhood of the central value, it is indifferent what distribution of *a priori* probabilities of p we suppose. Over the small, important central region the assumption that the *a priori* probability of p over a region is proportional to that region is likely to be a good first approximation, whatever the actual law. (Edgeworth, *Statistical Journal*, 1908, p. 387, and references there given.)

We are not, then, liable to any considerable error from the assumption that underlies this and similar investigations, that the important values of the quantities sought are *a priori* equally probable at any point of the range of values that affects the analysis.

We may now sum up the result of finding p by sample. The most probable value in the universe is the observed value p'. The probability of a deviation, as great as p_1, from the observed value is given approximately by the normal error function with standard deviation

$$\sqrt{\frac{p'(1 - p')}{n}}, \text{ or } \sqrt{\frac{p'(1 - p')}{n}\left(1 - \frac{n}{N}\right)}$$

where N is the number in the universe and $\dfrac{n}{N}$ is not negligible.

The precision of a measurement is measured by the reciprocal of the standard deviation of the errors to which it is liable.

General Method.

More generally let X' be any given function of n samples chosen at random from a universe where the (unknown) corresponding function is X, and let $X = X' + x$.

If we can show that the chance of obtaining the value X', when the value in the universe is X, is of the form $P_x = P_0 e^{-\frac{1}{2}\frac{x^2}{\sigma^2}}$, where P_0 is the maximum chance and is obtained when $X = X'$, σ is constant, and $x = X - X'$, then we can affirm with reasonable certainty that the sample gives evidence that the most probable value of the function in question is X', and that the chance of deviations from X' is given by the normal function with standard deviation σ. In the case above, p' is X', p is X, x is $z\sigma$ and $n\sigma^2 = p'(1 - p')$. For the more general case the process of inversion is not quite so direct.*

In order then to determine the precision of any measurement based on a sample, we have to take three steps, the first to find the standard deviation of the difference between the true value and the observed value, the second to find the chance that any assigned deviation would arise, the third to apply the principle of inverse probability.

Precision of the Arithmetic Average.

In Chapter III it was shown that if n quantities were selected at random and independently from a frequency group, which satisfied certain conditions, that the chance that the average of the n quantities differed from the average of the universe by as much as x was

---

* If all values of X are *a priori* equally probable the chance that the observations came from a universe when the value of X was within the limits $X' \pm x$ is $2\int_0^x P_x \, dx$, if x is small, and by inverse probability the chance that the value in the universe was within these limits is

$$2\int_0^x P_x \, dx \div \int_{-\infty}^{\infty} P_x \, dx = \frac{2}{\sigma\sqrt{2\pi}} \int_0^x e^{-\frac{1}{2}\frac{x^2}{\sigma^2}} \, dx.$$

$$2\int_x^\infty \frac{1}{\sigma_a\sqrt{2\pi}}\, e^{-\frac{x^2}{2\sigma_a^2}}dx, \text{ where } \sigma_a = \frac{\sigma}{\sqrt{n}},$$

σ being the standard deviation in the universe.

It will be shown immediately that σ', the observed standard deviation of the sample, differs from σ by a quantity commensurate with $\frac{\sigma}{\sqrt{n}}$, and hence if n is large, σ' may be taken as equivalent to σ.

We may now complete the argument and say that if in a sample of n things, drawn independently from a group in which no large portion is distant more than, say, twice its standard deviation from its average, the average of the sample is $\bar{x}$ and its standard deviation is σ, then the chance that the average in the universe differs from $\bar{x}$ by as much as x is

$$2\int_x^\infty \frac{\sqrt{n}}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\frac{x^2}{\left(\frac{\sigma}{\sqrt{n}}\right)^2}}\, dx, \quad . \quad . \quad . \quad . \quad . \quad (118)$$

when n is large.

Precision of the Standard Deviation.

We will now extend this theorem to test the precision of the standard deviation and second moment as determined from the sample.

Let $\bar{x}, \sigma, \mu_2 \ldots$ be the unknown constants of the universe, and $\bar{x} + \bar{x}', \sigma', \mu_2' \ldots$ be the corresponding values calculated from the sample.

Let $\bar{x} + x_t$ be any observation, and write $x_t' = x_t - \bar{x}'$.

The frequency curve of x_t^2 has μ_2 for its average, and its standard deviation is given by $\sigma_d^2 = $ mean $(x_t^2 - \mu_2)^2 = \mu_4 - \mu_2^2$.

Its fourth moment is mean $(x_t^2 - \mu_2)^4 = \mu_8 - 4\mu_6\mu_2 + 6\mu_4\mu_2^2 - 3\mu_2^4 = M_4$, say.

In the universe $\frac{\mu_s}{\sigma^s}$ is finite by hypothesis for all values of s, and

therefore $\frac{M_4}{\sigma_d^4} = \left(\frac{\mu_8}{\sigma^8} - 4\frac{\mu_6}{\sigma^6} + 6\frac{\mu_4}{\sigma^4} - 3\right) \div \left(\frac{\mu_4}{\sigma^4} - 1\right)^2$ is finite.

Similarly, for any moment of x_t^2, $M_s \div \sigma_d^s$ is finite.

Hence the average of the quantities, x_t^2, as occurring in the

sample, by the theorem summarised on p. 312 and just used, may differ from μ_2 by an error with normal frequency and standard deviation

$$\frac{\sigma_d}{\sqrt{n}} = \sqrt{\frac{\overline{\mu_4 - \mu_2{}^2}}{n}}.$$

This average m_2, say,

$$= \frac{Sx_t{}^2}{n} = \frac{1}{n} S(x'_t + \bar{x}')^2 = \frac{1}{n} Sx'_t{}^2 + \bar{x}'^2. \text{ since } Sx'_t = 0, = \mu_2' + \bar{x}'^2.$$

Now $\bar{x}'^2$ is of order $\frac{1}{n}$ from formula (118), and the error in m_2 has just be shown to be of order $\frac{1}{\sqrt{n}}$. Hence $\bar{x}'^2$ can be neglected, and μ_2' written for m_2.

Hence the observed μ_2' differs from μ_2 in the universe by an error with normal frequency and standard deviation

$$\sqrt{\frac{\overline{\mu_4 - \mu_2{}^2}}{n}} \quad \ldots \ldots \ldots \ldots \text{(119)}$$

But $\sigma^2 = \mu_2$, $\therefore \delta\sigma = \frac{\delta\mu_2}{2\sqrt{\mu_2}}$; hence σ' differs from σ by an error with normal frequency and standard deviation

$$\sqrt{\frac{\overline{\mu_4 - \mu_2{}^2}}{4n\mu_2}} \quad \ldots \ldots \ldots \ldots \text{(120)}$$

If the universe was normal, $\mu_4 = 3\mu_2{}^2$, and the standard deviations of the observed standard deviation and second moment become

$$\frac{\sigma}{\sqrt{2n}} \text{ and } \mu_2\sqrt{\frac{2}{n}} \quad \ldots \ldots \ldots \text{(121)}$$

respectively.

By a similar method the standard deviation of

$$m_4 = \frac{Sx_t{}^4}{n} \text{ is } \sqrt{\frac{\overline{\mu_8 - \mu_4{}^2}}{n}},$$

where $m_4 = \frac{1}{n} S(x_t' - \bar{x}')^4 = \mu_4' - 4\bar{x}'\mu_3' + 6\bar{x}'^2\mu_2' - 3\bar{x}'^4.$

Hence the error in μ_4' is of as low order as that in m_4, that is of order $\frac{1}{\sqrt{n}}$.

We may therefore, in calculating the standard deviations of μ_3 and σ, replace the unknown μ_4 and μ_2 by the known μ_4' and μ_2', and in calculating the standard deviation of the average we are justified in writing σ' for σ.

The standard deviations and frequency curve of errors in higher moments or in the correlation coefficient cannot be, at any rate readily, calculated by this method,* and the, whole basis is reconstituted and the arguments reset in the following paragraphs which are based on the papers to which reference is made at the head of the chapter.

Standard Deviations of the Average, etc., without Reference to Inverse Probability.

Suppose that in a universe containing N measurable objects, there are $N \times y_1$ at measurement x_1, $N \times y_2$ at $x_2 \ldots$, and that n objects are selected at random, n/N being so small that the chance of getting any value of x is not affected by previous selections.

$$N = N \times y_1 + N \times y_2 + \ldots, \quad \therefore y_1 + y_2 + \ldots = 1.$$

Let $\bar{x}'$, σ', and μ_2' be the average, standard deviation and second moment found from the samples. Required to determine the precision of these values as representatives of the average, standard deviation, and second moment for the universe.

Suppose x_1, $x_2 \ldots$ to be measured from the (unknown) average in the universe, so that $\mu_1 = S(x_t y_t) = 0$

Let μ_2 be the second moment for the universe, so that $\mu_2 = S . x_t^2 y_t$. and write $\mu_2 = \sigma^2$.

The sample will not, of course, contain precisely $n \times y_1$ at x_1, $n \times y_2$ at x_2 etc.

Let the numbers actually found be $n(y_1 + e_1)$ at $x_1 \ldots n(y_t + e_t)$ at $x_t \ldots$

Then $\qquad e_1 + e_2 + \ldots + e_t + \ldots = 0.$

x_1, $x_2 \ldots$ are, of course, constant.

Since y_t is the chance of finding an object at x_t and the experiment is made n times, e_t has normal frequency with standard deviation

$$\sqrt{\left\{\frac{y_t(1 - y_t)}{n}\right\}} \quad \text{(p. 278)}.$$

Hence the mean of all values of e_t^2 is $\dfrac{y_t(1 - y_t)}{n}$. and e_t is of order $1/\sqrt{n}$.

---

* The method is based on communications to the author from Professor Edgeworth.

Let E be the aggregate error in all the compartments together other than the sth and the tth. Write Y for $1 - y_s - y_t$.

Then

$$e_s + e_t + E = 0$$

$$\therefore 2e_s e_t = E^2 - e_s^2 - e_t^2$$

$$\therefore \text{Mean } e_s e_t = \tfrac{1}{2} \text{ mean } E^2 - \tfrac{1}{2} \text{ mean } e_s^2 - \tfrac{1}{2} \text{ mean } e_t^2$$

$$= \frac{1}{2n}\{Y(1 - Y) - y_s(1 - y_s) - y_t(1 - y_t)\}$$

$$= \frac{1}{2n}\{(1 - y_s - y_t)(y_s + y_t) - y_s(1 - y_s) - y_t(1 - y_t)\}$$

$$= -\frac{y_s y_t}{n}.$$

Now let F be any linear function of $y_1, y_2 \ldots$, so that

$$F = a_1 y_1 + a_2 y_2 + \ldots,$$

where $a_1\ a_2 \ldots$ are known constants.

Write $F + f = a_1(y_1 + e_1) + \ldots + a_t(y_t + e_t) + \ldots$,

so that

$$f = a_1 e_1 + \ldots + a_t e_t + \ldots$$

$$f^2 = S a_t^2 e_t^2 + 2S a_s a_t e_s e_t.$$

Then, if σ_f^2 is written for the mean value of f^2 when all possible values of $e_1, e_2, \ldots$ have been found in due proportion,

$$\sigma_f^2 = S a_t^2 \cdot (\text{mean } e_t^2) + 2S a_s a_t (\text{mean } e_s e_t)$$

$$= \frac{1}{n}\{S a_t^2 y_t (1 - y_t) - 2S a_s a_t y_s y_t\}$$

$$= \frac{1}{n}\{S a_t^2 y_t - F^2\} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (122)$$

Put

$$a_1 = x_1 \ldots a_t = x_t \ldots$$

$$F = S x_t y_t = 0$$

$$\bar{x}' = F + f = S x_t e_t$$

$$\therefore \sigma_{\bar{x}}^2 (= \text{mean of } \bar{x}'^2) = \frac{1}{n} S x_t^2 y_t \text{ from } (122) = \frac{\mu_2}{n} \quad . \quad . \quad . \quad (123)$$

$\bar{x}'$ is therefore of the order $\dfrac{1}{\sqrt{n}}$.

Now put $a_1 = x_1^2 \ldots\ a_t = x_t^2$

$$\mu_2 = F = S x_t^2 y_t, \quad \mu_2' = S(x_t - \bar{x}')^2 (y_t + e_t)$$

$$\mu_2' - \mu_2 = f = S x_t^2 e_t - 2\bar{x}' S x_t y_t + \text{terms involving } \bar{x}' e_t \text{ and } \bar{x}'^2$$

which are of order $\dfrac{1}{n}$.

$\therefore \mu_3' - \mu_3 = \mathrm{S}x_i{}^2 e_i$, since $\mathrm{S}x_i y_i = \dot{0}$, if terms in $\dfrac{1}{n}$ are neglected.

$\therefore$ from (122) $\sigma^2{}_{\mu_3} = $ mean of $(\mu_3' - \mu_3)^2$

$$= \frac{1}{n}\{\mathrm{S}(x_i{}^4 y_i) - \mu_3{}^2\} = \frac{\mu_4 - \mu_2{}^2}{n} \quad \ldots \quad (124)$$

Now $\sigma^2 = \mu_3$

Hence increments $\delta\sigma$, $\delta\mu$ of σ and μ_2 are connected by the equation

$$2\sigma\delta\sigma = \delta\mu_3, \text{ or } \delta\sigma = \frac{\delta\mu_3}{2\sqrt{\mu_3}} \quad \ldots \quad (125)$$

Hence $\qquad \sigma_\sigma{}^2 = \text{mean }(\delta\sigma)^2 = \text{mean } \dfrac{(\delta\mu_3)^2}{4\mu_3}.$

$\therefore \qquad\qquad\qquad \sigma_\sigma{}^2 = \dfrac{\mu_4 - \mu_2{}^2}{4\mu_3 n}$ from (124).

A similar analysis leads to the general result

$$\sigma_{\mu_p}{}^2 = \frac{1}{n}(\mu_{2p} - 2p\mu_{p+1}\mu_{p-1} + p^2\mu_{p-1}^2\mu_3 - \mu_p{}^2) \quad . \quad . \ (126)$$

Hence the standard deviations of $\bar{x}'$, σ and all moments involve

the factor $\dfrac{1}{\sqrt{n}}$, and if n is large the difference between the apparent

and true measurements is of the order $\dfrac{1}{\sqrt{n}}$ and may be neglected

in formulæ involving them. Consequently in evaluating the formulæ 123 to 126 the calculated values of the moments μ_3' etc. can be substituted for the unknown true values.

Notice that the standard deviation of each moment depends on the moment of twice its order, and this higher moment rapidly becomes great as the order is increased. In practice it is found that with ordinary values of n the moments above the 4th lack precision for this reason.

If the frequency curve of the universe is normal $\mu_4 = 3\mu_3{}^2$, and

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}, \quad \sigma_{\mu_2} = \sigma^2\sqrt{\frac{2}{n}}, \quad \sigma_\sigma = \frac{\sigma}{\sqrt{2n}},$$

$$\sigma_{\mu_3} = \sigma^3\sqrt{\frac{6}{n}}, \quad \sigma_{\mu_4} = 4\sigma^4\sqrt{\frac{6}{n}} \quad \ldots \quad (127)$$

The first two results for the normal curve can be obtained more directly as follows. Let $X_1, X_2 \ldots X_n$ be the measurements of n things taken at random from a group whose frequency curve is

$$y = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-x_0)^2}{2\sigma^2}},$$

where x_0 and σ are unknown.

P_x, the probability that these particular n things will be selected, is

$$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(X_1-x_0)^2}{2\sigma^2}} \times \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(X_2-x_0)^2}{2\sigma^2}} \times \ldots = \frac{1}{\sigma^n(2\pi)^{\frac{1}{2}}}e^{-S\frac{(X_i-x_0)^2}{2\sigma^2}}$$

Let $\bar{x}$ be the average of the X's, $X_i = \bar{x} + x_i$, etc., so that $Sx_i = 0$.

$$\log P_x = -n\log\sigma - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}S\{\bar{x} - x_0 + x_i\}^2$$

$$= -n\log\sigma - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}\{n(\bar{x} - x_0)^2 + ns^2\},$$

where s is the standard deviation of the X's.

Here $\bar{x}$ and s are known and σ and x_0 are to be determined.

P_x is greatest when $\bar{x} - x_0$ is least, whatever the value of σ.

Give x_0 the value $\bar{x}$

Then P_x is greatest when $\dfrac{dP_x}{d\sigma}$ is zero, that is when

$$0 = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\cdot ns^2, \text{ and } \sigma = s.$$

Write $\sigma = s + \gamma$ and $x_0 = \bar{x} + \delta$, and let P_0 be the value of P_x when $\gamma = 0 = \delta$.

Then $\log P_x - \log P_0 = -n\log\dfrac{\sigma}{s} - \dfrac{n}{2\sigma^2}(\delta^2 + s^2) + \dfrac{n}{2s^2}\cdot s^2$

$$\frac{1}{n}\log\frac{P_x}{P_0} = -\log\left(1 + \frac{\gamma}{s}\right) - \frac{1}{2s^2}\left(1 + \frac{\gamma}{s}\right)^{-2}(\delta^2 + s^2) + \frac{1}{2}$$

$$= -\frac{\gamma}{s} + \frac{\gamma^2}{2s^2} - \ldots - \frac{\delta^2}{2s^2} + \frac{2\gamma}{2s^2}\cdot s^2 - \frac{3\gamma^2}{2s^4}\cdot s^2, \text{ neglecting } \gamma^3, \gamma\delta^2 \text{ etc.},$$

$$= -\frac{\gamma^2}{s^2} - \frac{\delta^2}{2s^2}.$$

$$\therefore \qquad P_x = P_0\, e^{-\frac{1}{2}\cdot\frac{\delta^2}{\left(\frac{s}{\sqrt{n}}\right)^2}}\cdot e^{-\frac{1}{2}\frac{\gamma^2}{\left(\frac{s}{\sqrt{2n}}\right)^2}} \qquad \ldots \ldots \quad (128)$$

Hence the errors in $\bar{x}$ and s are independent of each other, are of normal frequency, and their standard deviations are respectively $\dfrac{s}{\sqrt{n}}$ and $\dfrac{s}{\sqrt{2n}}$, where s is the standard deviation of the sample, and for it σ, the standard deviation in the universe may be written without perceptible error when n is large, as in the results already obtained.

Standard Deviation of the Correlation Coefficient.

The standard deviation of the error to which the correlation coefficient is liable may be found as follows. (Here Dr. Sheppard's method is followed.)

Let there be pairs of values, such as (x_t, y_t), measured from their averages, whose standard deviations are σ_1, σ_2 and second moments λ, μ. Let the whole number of pairs be N, and let z_tN be situated at $(\bar{x} + x_t, \bar{y} + y_t)$, so that $z_1 + \ldots + z_t \ldots = 1$.

Also $S z_t x_t = 0 = S z_t y_t$.

Write M_{st} for the mean of $x^s y^t$ taken over all the pairs, so that $M_{st} = S z x^s y^t$. Write M for M_{11}.

Then if r is the coefficient of correlation of the N pairs,

$$r = \frac{M}{\sigma_1 \sigma_2} \quad \text{and} \quad \log r = \log M - \tfrac{1}{2} \log \lambda - \tfrac{1}{2} \log \mu.$$

Now let a selection of n pairs be made at random, and let the number selected from the position x_t, y_t be $(z_t + e_t)\, n$.

If x', y' are the resulting averages

$$x' = S(z_t + e_t) x_t = S x_t e_t, \quad \text{and} \quad y' = S y_t e_t.$$

The resulting deviations between the values in the sample and the values in the universe of r, M, λ, μ are, by differentiating the equation for $\log r$, evidently connected by

$$\frac{\delta r}{r} = \frac{\delta M}{M} - \frac{\delta \lambda}{2\lambda} - \frac{\delta \mu}{2\mu}.$$

Now

$$\delta M = S(z_t + e_t)(x_t - x')(y_t - y') - S z_t x_t y_t = S x_t y_t e_t - x' . S z_t y_t - y' . S z_t x_t,$$

when products of any two of the small quantities e_t, x', y', which are of order $\dfrac{1}{\sqrt{n}}$ are neglected.

$$\therefore \ \delta M = S x_t y_t e_t.$$

As shown above (pp. 419–20),

$$\delta \lambda = S x_t^2 e_t \quad \text{and} \quad \delta \mu = S y_t^2 e_t.$$

$$\therefore \qquad \frac{\delta r}{r} = S\left(\frac{x_t y_t}{M} - \frac{x_t^2}{2\lambda} - \frac{y_t^2}{2\mu}\right) e_t.$$

Hence, from the general formula (122), if σ_r is the standard deviation of the errors in r,

$$\sigma_r^2 = \frac{r^2}{n}\left[S\left(\frac{x_t y_t}{M} - \frac{x_t^2}{2\lambda} - \frac{y_t^2}{2\mu}\right)^2 z_t - S\left\{\left(\frac{x_t y_t}{M} - \frac{x_t^2}{2\lambda} - \frac{y_t^2}{2\mu}\right) z_t\right\}^2 \right]$$

$$= \frac{r^2}{n}\left\{ \frac{M_{22}}{M^2} + \frac{\lambda_4}{4\lambda^2} + \frac{\mu_4}{4\mu^2} - \frac{M_{31}}{\lambda M} - \frac{M_{13}}{\mu M} + \frac{M_{22}}{2\lambda\mu} \right\}.$$

since $\qquad S\left(\dfrac{x_t y_t}{M} - \dfrac{x_t^2}{2\lambda} - \dfrac{y_t^2}{2\mu}\right) z_t = I - \tfrac{1}{2} - \tfrac{1}{2} = 0,$

where λ_4 and μ_4 are the fourth moments of the distribution of the N x's and N y's.

In the case where the original distribution is that given by the normal correlation surface, $\lambda_4 = 3\lambda^2$, $\mu_4 = 3\mu^2$, $M = r\sigma_1\sigma_2$, $M_{22} = (1 + 2r^2)\,\sigma_1^2\sigma_2^2$, $M_{31} = 3r\sigma_1^3\sigma_2$, $M_{13} = 3r\sigma_1\sigma_2^3$ (formula (106)),

and $\quad \sigma_r^2 = \dfrac{r^2}{n}\left\{\dfrac{1 + 2r^2}{r^2} + \tfrac{3}{4} + \tfrac{3}{4} - 3 - 3 + \dfrac{1 + 2r^2}{2}\right\} = \dfrac{(1 - r^2)^2}{n},$

and $\qquad\qquad\qquad\qquad \sigma_r = \dfrac{1 - r^2}{\sqrt{n}}. \quad . \quad . \quad . \quad . \quad . \quad . \quad (129)$

This is the value generally used, it being implicitly assumed that the distribution approximates to the normal.

The regression coefficient, when y is expressed in terms of x,

is $r\dfrac{\sigma_y}{\sigma_x} = \rho$, say. In the present notation $\rho = \dfrac{M}{\lambda}$, and we obtain by a method similar to that just used,

$$\sigma_\rho^2 = \dfrac{\rho^2}{n}\left\{\dfrac{M_{22}}{M^2} + \dfrac{\lambda_4}{\lambda^2} - \dfrac{2M_{31}}{\lambda M}\right\} \text{ in any distribution.}$$

Hence in normal distribution

$$\sigma_\rho^2 = \dfrac{\rho^2}{n}\left\{\dfrac{1 + 2r^2}{r^2} + 3 - 6\right\} = \dfrac{\rho^2}{n} \cdot \dfrac{1 - r^2}{r^2}$$

$$\therefore \sigma_\rho = \dfrac{1}{\sqrt{n}} \cdot \dfrac{\sigma_y}{\sigma_x} \cdot \sqrt{1 - r^2}.$$

In the case of normal distribution the result may be reached as follows. (Here Professor Pearson's method is followed.)

Suppose pairs $x_1, y_1 \ldots x_n, y_n$ are chosen from a surface whose unknown centre is x_0, y_0, standard deviations σ_1, σ_2, and mean product $r\sigma_1\sigma_2$.

Let $\bar{x}, \bar{y}, s_1, s_2, r'$ be calculated from the sample.

The chance of concurrence of the n pairs is

$$P_s = \dfrac{1}{(2\pi\sigma_1\sigma_2\sqrt{1 - r^2})^n}\, e^{-\frac{1}{2(1 - r^2)} \cdot S\left\{\frac{(x_t - x_0)^2}{\sigma_1^2} + \frac{(v_t - y_0)^2}{\sigma_2^2} - \frac{2r(x_t - x_0)(v_t - y_0)}{\sigma_1\sigma_2}\right\}}.$$

$$\therefore \log P_s = - n \log 2\pi\sigma_1\sigma_2 - \dfrac{n}{2}\log(1 - r^2)$$

$$- \dfrac{n}{2(1 - r^2)}\left\{\dfrac{s_1^2 + d_1^2}{\sigma_1^2} + \dfrac{s_2^2 + d_2^2}{\sigma_2^2} - \dfrac{2r \cdot (r's_1s_2 + d_1d_2)}{\sigma_1\sigma_2}\right\}$$

where $d_1 = x_0 - \bar{x}$, $d_2 = y_0 - \bar{y}$.

Here $r, \sigma_1, \sigma_2, d_1, d_2$ are unknown, and r', s_1, s_2 known.

By expressing the conditions

$$\frac{\partial P}{\partial d_1} = o = \frac{\partial P}{\partial d_2} = \frac{\partial P}{\partial \sigma_1} = \frac{\partial P}{\partial \sigma_2} = \frac{\partial P}{\partial r},$$

we obtain the values of the five unknowns which make P_s a maximum.

$$\frac{d_1}{\sigma_1^2} - \frac{rd_2}{\sigma_1\sigma_2} = o = \frac{d_2}{\sigma_2^2} - \frac{rd_1}{\sigma_1\sigma_2},$$

whence $d_1 = d_2 = o$, unless $r^2 = 1$.

Then, taking d_1 and d_2 to be zero,

$$\frac{1}{\sigma_1} - \frac{s_1^2}{(1-r^2)\sigma_1^2} + \frac{rr's_1s_2}{(1-r^2)\sigma_1^2\sigma_2} = o = \frac{1}{\sigma_2} - \frac{s_2^2}{(1-r^2)\sigma_2^2} + \frac{rr's_1s_2}{(1-r^2)\sigma_1\sigma_2^2}$$

$$\therefore \qquad (1-r^2)\sigma_1^2\sigma_2 = s_1^2\sigma_2 - rr's_1s_2\sigma_1$$

and $\qquad (1-r^2)\sigma_1\sigma_2^2 = s_2^2\sigma_1 - rr's_1s_2\sigma_2,$

whence $\qquad \dfrac{s_1}{\sigma_1} = \dfrac{s_2}{\sigma_2} = k$, say, and $\quad 1 - r^2 = k^2(1 - rr')$

and $\qquad \dfrac{r}{1-r^2} - \dfrac{r}{(1-r^2)^2}\left\{\dfrac{s_1^2}{\sigma_1^2} + \dfrac{s_2^2}{\sigma_2^2} - \dfrac{2rr's_1s_2}{\sigma_1\sigma_2}\right\} + \dfrac{r's_1s_2}{(1-r^2)\sigma_1\sigma_2} = o.$

$$\therefore \qquad r(1-r^2) - 2rk^2(1-rr') + r'(1-r^2)k^2 = o$$

Hence $r = r'$ and $k = 1$.

P_s is greatest, therefore, when the values found in the sample are taken as the values in the surface. Write P_0 for the value of P_s so obtained.

Now write $\sigma_1 = s_1 + \gamma_1$, $\sigma_2 = s_2 + \gamma_2$, and $r = r' + \rho$, and expand all functions in powers of the small quantities d_1, d_2, γ_1, γ_2, ρ, neglecting third powers.

We obtain

$$\frac{1}{n}\log\frac{P_s}{P_0} = -\frac{1}{2(1-r'^2)}\left(\frac{d_1^2}{s_1^2} + \frac{d_2^2}{s_2^2} - \frac{2r'd_1d_2}{s_1s_2}\right) + \frac{r'}{1-r'^2}\left(\frac{\rho\gamma_1}{s_1} + \frac{\rho\gamma_2}{s_2}\right)$$

$$+ \frac{r'^2}{1-r'^2}\cdot\frac{\gamma_1\gamma_2}{s_1s_2} - \frac{2-r'^2}{2(1-r'^2)}\left(\frac{\gamma_1^2}{s_1^2} + \frac{\gamma_2^2}{s_2^2}\right) - \frac{1+r'^2}{2(1-r'^2)^2}\rho^2$$

$$= -\frac{1}{2(1-r'^2)}\left(\frac{d_1}{s_1} - \frac{r'd_2}{s_2}\right)^2 - \frac{d_2^2}{2s_2^2} - \frac{2-r'^2}{2(1-r'^2)}\left(\frac{\gamma_1}{s_1} - \frac{r'^2}{2-r'^2}\cdot\frac{\gamma_2}{s_2} - \frac{r'}{2-r'^2}\rho\right)^2$$

$$- \frac{2}{2-r'^2}\left(\frac{\gamma_2}{s_2} - \frac{r'\rho}{2(1-r'^2)}\right)^2 - \frac{\rho^2}{2(1-r'^2)^2}.$$

Integrate successively between extreme limits

for d_1, regarding d_2, γ_1, γ_2, ρ as constant,

for d_2, regarding γ_1, γ_2, ρ as constant,

for γ_1, regarding γ_2, ρ as constant,

and for γ_2, regarding ρ as constant.

We then find that the whole chance of the observations arising from a value $r' + \rho$, whatever the values of x_0, y_0, σ_1, σ_2 is

$$P = Ke^{-\frac{1}{2}\left(\frac{\rho^2}{1-r'^2}\right)^2}.$$

That is, the distribution is normal, with standard deviation of $r' = \dfrac{1-r^2}{\sqrt{n}}$.

The work on pp. 421 and 424 shows that if the frequency group from which samples are taken is normal, then the chances of obtaining various errors in $\bar{x}$, σ, μ and r in the sample are given by the normal probability function ; and inversely, that the chances that the corresponding quantities in the universe have various deviations from the observed quantities are also so given. It remains to prove that under other conditions the same result is obtained.

In each case the quantity concerned was put in the form

$$F + f = a_1(y_1 + e_1) + a_2(y_2 + e_2) + \ldots,$$

where $e_1 + e_2 + \ldots = 0$, and $f = 0$ when $0 = e_1 = e_2 = \ldots$. Also $y_1 + y_2 + \ldots = 1$.

The frequency curves of e_1, e_2 ... are normal if n, the number in the sample, is large, p. 418.

If e_1, e_2 ... were independent of each other, or if the number of separate values of x_1, x_2 ... were so great that we could treat them as independent, then we could at once apply the theorem of pp. 295 *seq.* and state the frequency of f is normal.

The full analysis (given in Appendix, Note 9) leads to the result that normality may be presumed under the same conditions affecting the universe from which the samples are taken as lead to normality of the average, viz. : that the universe is so confined that the ratio $\frac{\mu_t}{\sigma^i}$ is finite for all values of t. (p. 299).

Note added in 1936.—It should perhaps have been more explicitly stated that the methods of pp. 421–4 do not relate directly to the same problem as those of pp. 418–20. The last-named supposes many samples from one universe, the first considers the probability (or likelihood) of a given example from various universes. The *forms* of the results tend to be the same as n increases ; but the treatment of the equations (128) and at the bottom of p. 424 as frequency groups involves *à priori* probability, as indicated at the top of this page. The problem " given the target how will shots be dispersed " leads to Prof. R. A. Fisher's method of " variance " ; the problem " given the shot-marks, what was the target," which is the practical question in many cases when we have only one sample, necessarily involves the reference to *à priori* probability.

CHAPTER X.

TESTS OF CORRESPONDENCE BETWEEN DATA* AND FORMULÆ.

In the general method of the representation of observations by a mathematical formula, the question must arise how the adequacy of the formula is to be tested, or, as it is frequently phrased, a test of the goodness of fit is required.

Consider for example the table used above (p. 310) of the weekly expenditure on food per " unit " in 970 families.

Expenditure.	m' number of cases.	m calculated numbers.	$\epsilon = m \sim m'$ difference.	Standard deviations.	$\dfrac{\epsilon^2}{m}$.
Not exceeding 5·5s..	18	22	4	4·6	·7
5·5	107	123	16	10·4	2·1
7·5	255	234	21	13·3	1·9
9·5	245	249	4	13·6	·1
11·5	173	168	5	11·8	·1
13·5	101	89	12	9·0	1·6
15·5	38	51	13	7·0	3·3
17·5	17	22	5	4·6	1·1
19·5	9	11	2	3·3	·4
Over 21·5 . . .	7	1	6	?	36·0
Totals . .	970	970	88	—	47·3

The calculated numbers are from the second approximation to the Law of Great Numbers. A rough method formerly used was to add the differences between the calculated numbers and the numbers observed in each compartment, irrespective of sign, and to express this total as a percentage of the number of cases. The " percentage misfit " thus calculated is $88 \div 9\cdot70 = 9\cdot1$ per cent.

The weakness of this method is that it is not related to any measurement of probability, and one cannot tell at sight whether the fit is good or not. Of two competing formulæ, the presumption is that that which gives the lower percentage misfit is the better ; also when we have several sets of similar

426

observations we can tell roughly by this method which is nearest to the formula, and in some cases in which set the observations are most regular.

The percentage misfit is generally diminished if compartments are merged together.

As regards the contents of individual compartments, we already have a simple test. If m_t is the calculated number in a compartment when there are N observations in all, the chance of finding $m_t + e_t$ observations in this compartment in a random selection is

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2} \cdot \frac{e_t^2}{\sigma^2}} \text{ (formula (19)) where } \sigma^2 = \frac{m_t}{N}\left(1 - \frac{m_t}{N}\right)N,$$

and the probability of exceeding any assigned multiple or sub-multiple of σ is given by the table (p. 271). The standard deviation for each grade in the above example except the last is given, and it is seen that four out of nine errors are less than σ, their standard deviation, two are between σ and $\frac{3\sigma}{2}$, and the remaining three less than 2σ. No separate measurement is improbable, and therefore the whole grouping may be presumed to be not improbable, except the final number, 7 above 21·5s.

That numbers in extreme grades should be discontinuous in relation to middle grades is common in many classes of observations.

The deviations are not independent, however, since their total must be zero ; and even if the deviation in one compartment taken by itself is improbably large, it may yet not be improbable when all the compartments are considered. A measurement which allows for this modification has been devised by Professor Pearson, and part of the analysis in a simplified form, a brief table of the results, and some applications are given in the following paragraphs (see *The Philosophical Magazine*, No. 302, July, 1900, pp. 157–175).

Suppose that a formula, which is presumed to represent the distribution of observations, leads to the expectation of $m_1, m_2 \ldots m_n$ observations in n grades or compartments, when N, $= m_1 + m_2 + \ldots + m_n$, is the whole number of observations.

In an experiment or group of observations, suppose that

$(m_1 + e_1) \ldots (m_t + e_t) \ldots (m_n + e_n)$ are found in the compartments, so that $e_1 + \ldots + e_t + \ldots + e_n = 0$.

Write $p_1 = \dfrac{m_1}{N} \ldots p_t = \dfrac{m_t}{N} \ldots$

Then p_t is the chance that an observation from a group satis-. fying perfectly the formula will fall into the tth grade.

The chance that $m_t + e_t$ will fall into this grade when N are chosen at random from an indefinitely large universe is

$$\frac{1}{\sigma_t\sqrt{2\pi}} e^{-\frac{e_t^2}{2\sigma_t^2}},$$

where $\quad \sigma_t^2 = p_t(1 - p_t)N = p_tq_tN$, where $\quad q_t = 1 - p_t$.

It can be shown that the joint chance of the errors named is

$$Ke^{-\frac{1}{2}x^2}, \quad \text{where } x^2 = S \cdot \frac{e_t^2}{m_t}, \quad \text{and} \quad Se_t = 0,$$

K being a constant.

For, if there were only *two* compartments, $e_1 + e_2 = 0$, and the joint chance equals the chance of either.

Then $\qquad p = \dfrac{m_1}{N}, \quad q = \dfrac{m_2}{N}, \quad m_1 + m_2 = N.$

The chance is

$$\frac{N^{\frac{1}{2}}}{\sqrt{2\pi m_1 m_2}} e^{-\frac{1}{2}\left(\frac{e_1^2}{m_1} + \frac{e_2^2}{m_2}\right)}, \text{ since } \frac{e_1^2 N}{m_1 m_2} = \frac{e_1^2(m_2 + m_1)}{m_1 m_2}, \text{ and } e_1^2 = e_2^2.$$

If there are *three* compartments

$$e_1 + e_2 + e_3 = 0, \quad m_1 + m_2 + m_3 = N, \quad \sigma_1^2 = \frac{m_1}{N} \cdot \frac{m_2 + m_3}{N} \cdot N,$$

and similarly for σ_2^2 and σ_3^2.

$$2e_1e_2 = e_3^2 - e_1^2 - e_2^2,$$

$$r\sigma_1\sigma_2 = \text{mean } e_1e_2 = \tfrac{1}{2}(\sigma_3^2 - \sigma_1^2 - \sigma_2^2)$$

$$= \frac{1}{2N}\{m_3(m_1 + m_2) - m_1(m_2 + m_3) - m_2(m_1 + m_3)\}$$

$$= -\frac{m_1 m_2}{N}. \quad \text{(Compare p. 419.)}$$

The chance of the concurrence of e_1 and e_2, and therefore of e_3 also, is given by the normal correlation surface as

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1 - r^2}} e^{-\frac{1}{2(1-r^2)}\left(\frac{e_1^2}{\sigma_1^2} + \frac{e_2^2}{\sigma_2^2} - \frac{2re_1e_2}{\sigma_1\sigma_2}\right)}$$

Now

$$\sigma_1{}^2\sigma_3{}^2(1-r^2) = \frac{m_1 m_2(m_2 + m_3)(m_1 + m_3)}{N^2} - \frac{m_1{}^2 m_2{}^2}{N^2} = \frac{m_1 m_2 m_3}{N},$$

since $m_1 + m_2 + m_3 = N$.

Hence the index of e is

$$-\frac{N}{2m_1 m_2 m_3}(e_1{}^2\sigma_2{}^2 + e_2{}^2\sigma_1{}^2 - 2r\sigma_1\sigma_2 e_1 e_2)$$

$$= -\frac{N}{2m_1 m_2 m_3}\left\{ \frac{e_1{}^2 m_2(m_1 + m_3)}{N} + \frac{e_2{}^2 m_1(m_2 + m_3)}{N} + \frac{2e_1 e_2 m_1 m_3}{N} \right\}$$

$$= -\frac{1}{2m_1 m_2 m_3}\{(e_1 + e_2)^2 m_1 m_2 + e_1{}^2 m_2 m_3 + e_2{}^2 m_1 m_3\}$$

$$= -\tfrac{1}{2}\left(\frac{e_1{}^2}{m_1} + \frac{e_2{}^2}{m_2} + \frac{e_3{}^2}{m_3} \right), \text{ since } e_1 + e_2 = -e_3.$$

Now if the second and third compartments had been merged into one containing $M + E$ observations, where $M = m_2 + m_3$ and $E = e_2 + e_3$, the chance would have been

$$K_1 . e^{-\frac{1}{2}\left(\frac{e_1{}^2}{m_1} + \frac{E^2}{M} \right)},$$

where K_1 is a constant.

The effect, therefore, of dividing the second compartment without changing the first is to alter the constant and to replace $\frac{E^2}{M}$ by $\frac{e_2{}^2}{m_2} + \frac{e_3{}^2}{m_3}$ in the index.

Similarly if two compartments are given, the effect of dividing the third compartment without changing the first two must be to alter the constant and to replace $\frac{e_3{}^2}{m_3}$ by $\frac{e_3{}^2}{m_3} + \frac{e_4{}^2}{m_4}$ in the index, and so on.

Hence for n compartments the chance, P, of errors $e_1, e_2 \ldots e_n$ is

$$Ke^{-\frac{1}{2}x^2}, \text{ where } x^2 = \frac{e_1{}^2}{m_1} + \frac{e_2{}^2}{m_2} + \ldots + \frac{e_n{}^2}{m_n},$$

and $$e_1 + e_2 + \ldots e_n = 0 \quad . \quad . \quad . \quad . \quad . \quad . \quad (130)$$

Notice that x^2 is the same expression as is used in obtaining the coefficient of contingency.

[A proof of the formula, without the above method of induction, is given by Pearson, by the use of the multiple correlation equation.   See also Note 11, p. 454 below.]

If the selections in the compartments had been independent

and without the condition that $e_1 + e_2 + \ldots = 0$, the chance would have been

$$Ke^{-\frac{1}{2}x^2} \times e^{-\frac{1}{2}\left(\frac{e_1^2}{N-m_1} + \frac{e_2^2}{N-m_2} + \ldots\right)}$$

for the index would have been

$$-\frac{1}{2}\left(\frac{e_1^2 N}{m_1(N-m_1)} + \ldots\right) = -\frac{1}{2}\left(\frac{e_1^2}{m_1} + \frac{e_1^2}{N-m_1} + \ldots\right).$$

If there are many compartments and the largest of the fractions $\frac{m_i}{N}$ is small, the second part of the index is negligible compared with the first, and the two expressions tend to equality, and the effect of the correlation is small.

The chance of the occurrences if there is no correlation is less than that when there is correlation, since the last factor, if not negligible, is less than 1. (The constant is eliminated in further processes.) Hence the aggregation of uncorrelated chances, which is simpler than the present method, gives an unduly unfavourable view of the appropriateness of a formula.

The chance of every system of errors that gives a particular value of x^2 is the same. Now, when the probability of a deviation from the mean in normal frequency is in question, it is customary to measure the probability that so great a deviation to left or right should have occurred, viz.,

$$2\int_z^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz.$$

Similarly here we may measure the chance of the occurrence of the system of errors or a less probable system by evaluating $2\iint \ldots Ke^{-\frac{1}{2}x^2} d_x$, where d_x is written for $de_1 . de_2 \ldots de_{n-1}$ and the integral is $\overline{n-1}$ fold and extended from x to ∞, with the condition $e_1 + e_2 + \ldots + e_n = 0$, K being so chosen that $\int_{-\infty}^\infty Ke^{-\frac{1}{2}x^2} d_x = 1$.

The existence of this condition makes the integration complicated, and reference should be made to Pearson's original analysis for its working out.

The result is that

$$P = \sqrt{\frac{2}{\pi}}\int_x^\infty e^{-\frac{1}{2}x^2} . d_x + \sqrt{\frac{2}{\pi}}e^{-\frac{1}{2}x^2}\left(\frac{x}{1} + \frac{x^3}{1\cdot3} + \ldots + \frac{x^{n-3}}{1\cdot3\ \overline{5-n-3}}\right)$$

when n is even, and

$$P = e^{-\frac{1}{2}x^2}\left(1 + \frac{x^2}{2} + \ldots + \frac{x^{n-3}}{2 \cdot 4 \ldots n-3}\right) \text{ when } n \text{ is odd. (131)}$$

A table of the values of P for various values of x^2 and n is given in *Biometrika*, Vol. I, pp. 155 *seq.* We can, in a very brief form, obtain a working rule for determining whether a formula does or does not adequately represent an observed group by picking out values of x^2 which for a given n make P = ½ or slightly more, or, further up the scale of improbability, make P = ·0455 or slightly less, which corresponds to twice the standard deviation in the normal curve.

n.	x^2.	P.	x^2.	P.
3	1	·61	6	·050
4	2	·57	8	·046
5	3	·56	10	·040
6	4	·55	12	·035
7	5	·54	13	·043
8	6	·54	15	·036
9	7	·54	16	·042
10	8	·53	18	·035
11	9	·53	19	·040
12	10	·53	20	·045
13	11	·53	22	·038
14	12	·53	23	·042
15	13	·53	24	·046
16	14	·526	26	·038
17	15	·525	27	·041
18	16	·524	28	·045
19	17	·523	30	·037
20	18	·522		
25	23	·520		
30	28	·518		

If $x^2 < n - 2$, it is at least an even chance—as likely as not—that the observations would be found from a group represented by the formula.

If $x^2 > 2n$, the improbability is considerable.

Strictly, the test should be applied using as many compartments as are given by the observations, for the merging of compartments affects the resulting value of P ; but it is often difficult to get back to ungraded observations, and in the case of continuous variables, such as height, the original grades would be as fine as the measurements could be made.

A more serious difficulty is that in any compartment the observed $m_t + e_t$ must be integral, while m_t is in general not integral, and some value of e_t would be found in the most perfect representation. In consequence, the number to be expected in the least occupied compartment must be reasonably

large, °or we obtain spurious contributions to x^2.   This in practice rules out detailed extreme compartments, and in their rejection or fusion an element of arbitrariness is introduced and no fine measurement is possible.

On the other hand, when we are testing the applicability, of the normal curve of error, or the general law of great numbers, based on Edgeworth's hypothesis (pp. 298–9), there is no expectation of closeness of fit on abscissæ beyond a small multiple of the standard deviation—the smaller as the number of independent elements that contribute to the measurement diminishes—so that the test is only applicable to the well-occupied central compartments ; but in choosing the extent over which the test is made, the fineness of the method is lost.

Hence, only a broad, but often sufficiently definite, result can be obtained.

Illustrations.

If we neglect the extreme grade in Example 7, on p. 310, $x^2 = 11\cdot3$, $n = 9$, P = ·18, and the formula " 2nd approx." is adequate.

If we take the Pearsonian formula, on the same page, $x^2 = 21\cdot4$, $n = 9$, P = ·006, but if we exclude the lowest as well as the highest grade, $x^2 = 4\cdot1$, $n = 8$, P = ·77 ; hence this formula expresses the central eight grades but not either extreme.

The same conclusions are reached if we simply take the standard deviations of the grades separately.

In the table on p. 309 relating to the ages of school children, $n = 8$.   The normal curve gives $x^2 = 16\cdot7$ and P = ·02, which is not satisfactory.   The second approximation, however, gives $x^2 = \cdot47$ and P is indistinguishable from 1.

In the experiment on the numbers of letters in words (pp. 305–6), the sum of 10 words, graded by 5 letters, gives $n = 13$, and with the normal curve $x^2 = 33$, P = ·001, or omitting the lowest and two highest extreme grades, $n = 10$, $x^2 = 6\cdot1$, P = ·73.   The second approximation, however, including all grades, gives $x^2 = 8\cdot4$, P = ·74.

The sums of 100 words graded by 20 letters give· $n = 10$, $x^2 = 2\cdot96$, P = ·965 with the normal curve, and no further approximation can improve on this.

An example of a different kind is found, when a distribution found by sample is compared with the whole group from which the sample is taken, to verify the rules of sampling or the adequacy of the method.

NUMBER OF COMPANIES PAYING DIVIDENDS AT VARIOUS RATES.

	Number in sample m'.	Relative numbers in all companies m.	Standard deviation.	$\frac{d^2}{m}$.
Below 3 per cent. . . .	34	30	5·3	·53
3 per cent.	108	108·8	8·9	0
4 ,, ...	117	124·4	9·3	·44
5 ,, ...	60	70·8	7·4	1·65
6 per cent. to 8 per cent. .	48	43·2	6·2	·53
8 per cent.	33	22·8	4·6	4·57
	400	400		7·72

Here $n = 6$, $X^2 = 7·72$, $P = ·185$ The result is fairly good but spoilt by the highest grade.

This test has been applied to the distribution in two dimensions, in the experiment tabulated on p. 394.

The 24 squares, 3 to left and right of centre, and 2 above and below it, which contain in theory 11 or more observations, were taken as separate compartments. Outlying squares were grouped in the 9 regions shown by the thick lines, rather arbitrarily, so as to get contiguous squares which aggregated to at least 9 expected observations in the second approximation. The results are as follows :—

	Normal surface.		2nd approximation.	
	X^2.	P.	X^2.	P.
24 central squares . . .	20·8	·59	17·5	·79
9 outlying regions . . .	27·8		10·1	
33 regions	48·6	·035	27·6	·59

The improvement in the outlying regions by the use of the second approximation is very marked.

Note.—In application of the test to double or manifold tables, as those on pp. 372-3, the procedure is different according as the sub-totals $n_1 \ldots m_1 \ldots$ are supposed to be given or not. See *Economica*, No. 7, p. 1, and No. 8, p. 139, and the *Statistical Journal*, 1922, pp. 87 seq.

APPENDIX.

MATHEMATICAL NOTES.

1.—*Wallis's Theorem for the Value of π.*

By simple graphic considerations it is evident that when n is a positive integer

$$\int_0^{\frac{\pi}{2}} \sin^{2n+1}x \,.\, dx < \int_0^{\frac{\pi}{2}} \sin^{2n}x \,.\, dx < \int_0^{\frac{\pi}{2}} \sin^{2n-1}x \,.\, dx$$

$$\therefore \quad \frac{2 \,.\, 4 \,.\, 6 \ldots 2n}{3 \,.\, 5 \,.\, 7 \ldots (2n+1)} < \frac{1 \,.\, 3 \,.\, 5 \ldots (2n-1)}{2 \,.\, 4 \,.\, 6 \ldots 2n} \,.\, \frac{\pi}{2} < \frac{2 \,.\, 4 \,.\, 6 \ldots (2n-2)}{3 \,.\, 5 \,.\, 7 \ldots (2n-1)} \,.$$

$$\therefore \quad \frac{2^{2n}(n!)^2}{(2n+1)!} < \frac{(2n)!}{2^{2n}(n!)^2} \frac{\pi}{2} < \frac{2^{2n}(n!)^2}{(2n)!} \,.\, \frac{1}{2n} \,.$$

$$\therefore \quad \frac{2^{2n}(n!)^2}{(2n)!\sqrt{2n+1}} < \sqrt{\frac{\pi}{2}} < \frac{2^{2n}(n!)^2}{(2n)!\sqrt{2n}} \,.$$

$$\therefore \quad \sqrt{\frac{\pi}{2}} = \frac{2^{2n}(n!)^2}{(2n)!\sqrt{2n}}, \text{ correct to } \frac{1}{n} \quad \ldots \ldots \ldots (132)$$

(See Gibson, *Treatise on the Calculus,* 1896, Ex. XXVI. 22.)

2.—*Sum of Powers of Integers.*

If we suppose

$$S_m = \sum_{t=1}^{t=m} t^r = am^{r+1} + bm^r + cm^{r-1} + \ldots,$$

we can find $a, b, c \ldots$ by induction.

For $(m + 1)^r = S_{m+1} - S_m = a\{(m+1)^{r+1} - m^{r+1}\}$
$$+ b\{(m+1)^r - m^r\} + c\{(m+1)^{r-1} - m^{r-1}\} \ldots$$

434

Equating coefficients of m^r, m^{r-1}, ..., we have

$$1 = a(r+1)$$

$$r = a\frac{(r+1)r}{2} + br, \text{ and } b = \tfrac{1}{2}$$

$$\frac{r(r-1)}{2} = a\frac{(r+1)r(r-1)}{6} + b\frac{r(r-1)}{2} + c(r-1), \text{ and } c = \frac{r}{12} \text{ etc.}$$

$$\therefore \frac{1^r + 2^r + \quad + m^r}{m^{r+1}} = \frac{1}{r+1} + \frac{1}{2m} + \frac{r}{12m^2} +$$

$$= \frac{1}{r+1}, \text{ if } \frac{1}{m} \text{ is neglected,}$$

$$= \frac{1}{r+1} + \frac{1}{2m}, \text{ if } \frac{1}{m^2} \text{ is neglected } . \quad . \quad . \quad . \ (133)$$

3.—Stirling's Formula for m !

The first approximation to this formula may be obtained from Wallis's Theorem as follows.

Write

$$z = \frac{(2m)\,!}{(2m)^{m+1}(m-1)\,!} = (2m-1)(2m-2)\ldots(2m-m) + (2m)^m.$$

Then

$$\log z = \sum_{t=1}^{t=m} \log\left(1 - \frac{t}{2m}\right)$$

$$-\log z = \frac{1+2+\quad+m}{2m} + \ldots + \frac{1^r + 2^r + \quad + m^r}{r(2m)^r} + \ldots$$

$$= \sum_{r=1}^{r=\infty} \left\{ \frac{1}{r\cdot 2^r}\cdot\left(\frac{m}{r+1} + \tfrac{1}{2} + \frac{r}{12m}\right) \right\},$$

by Note 2, if higher powers of $\frac{1}{m}$ are neglected.

Now $\displaystyle\sum \frac{2^{-r}}{r(r+1)} = \sum\frac{2^{-r}}{r} - 2\sum\frac{2^{-r-1}}{r+1}$

$$= -\log(1-\tfrac{1}{2}) + 2\{\log(1-\tfrac{1}{2}) + \tfrac{1}{2}\} = 1 - \log 2 = \log\left(\frac{e}{2}\right).$$

Hence

$$-\log z = m\log\left(\frac{e}{2}\right) - \tfrac{1}{2}\log(1-\tfrac{1}{2}) + \frac{1}{12m}$$

and $\quad z = \left(\frac{2}{e}\right)^m \cdot 2^{-\frac{1}{2}} \cdot e^{-\frac{1}{12m}} = 2^{m-\frac{1}{2}}\cdot e^{-m}\cdot\left(1 - \frac{1}{12m} + \ldots\right)$

$$= 2^{m-\frac{1}{2}}\cdot e^{-m}, \text{ if } \frac{1}{m} \text{ is neglected.}$$

But by Wallis's theorem

$$m! = \frac{(2m)!\,(2m)^{\frac{1}{2}}}{2^{2m}\cdot m!}\cdot\left(\frac{\pi}{2}\right)^{\frac{1}{2}},\text{ correct to }\frac{1}{m},\,\, = z \times \frac{(m\pi)^{\frac{1}{2}}\cdot m^{m}}{2^{m-1^{\frac{1}{2}}}}.$$

$$\therefore\ m! = m^{m}\cdot\sqrt{2\pi m}\cdot e^{-m}\ \cdots\cdots\cdots\cdots\ (134)$$

This formula gives an error of less than 1 per cent. for the value of 10! and rapidly reaches considerable accuracy if m is increased.

In its more complete form it is

$$m! = m^{m}\cdot\sqrt{2\pi m}\cdot e^{-m+\frac{1}{12m}-\frac{1}{360m^{3}}+}$$

(See Chrystal's *Algebra*, Chap. XXX.)

4.—*The Euler-Maclaurin Theorem, which connects Summation with Integration.*

Let $f(a)$, $f(a+h)$, ... $f(a+mh)$ be values of $f(x)$ at $\overline{m+1}$ successive values of x.

Then by Taylor's expansion

$$f(a+h) = f(a) + hf'(a) + \frac{h^{2}}{2}f''(a) + \ldots$$

$$f(a+2h) = f(a+h) + hf'(a+h) + \frac{h^{2}}{2}f''(a+h) + \ldots$$

. . . .
. . . .

$$f(a+mh) = f(a+\overline{m-1}h) + hf'(a+\overline{m-1}h) + \frac{h^{2}}{2}f''(a+\overline{m-1}h) + \ldots$$

Write $d = a + \overline{m-1}h$, and $b = a + mh$, and add.

$$f(b) - f(a) = h\sum_{a}^{d}F(x) + \frac{h^{2}}{2}\cdot\sum_{a}^{d}F'(x) + \frac{h^{3}}{3!}\sum_{a}^{d}F''(x) + \ldots,$$

where $F(x) = f'(x)$, and $\int F(x)\,.\,dx = f(x) + \text{constant}$.

$$\therefore\ h\sum_{a}^{d}F(x) = \int_{a}^{b}F(x)\,.\,dx - \frac{h^{2}}{2}\sum_{a}^{d}F'(x) - \frac{h^{3}}{3!}\sum_{a}^{d}F''(x) - \ldots$$

Similarly

$$h\sum_{a}^{d}F'(x) = \int_{a}^{b}F'(x)\,dx - \frac{h^{2}}{2}\sum_{a}^{d}F''(x) - \ldots$$

and $\quad h\sum_{a}^{d}F''(x) = \int_{a}^{b}F''(x)\,dx - \ldots$

Combining these equations, we have

$$h \sum_{a}^{d} F(x) = \int_{a}^{b} F(x) . dx - \frac{h}{2} \{F(b) - F(a)\} + \frac{h^2}{12} \{F'(b) - F'(a)\}$$

$$+ \text{ terms involving } h^4,$$

and ∴.

$$h \sum_{a}^{b} F(x) = \int_{a}^{b} F(x) . dx + \frac{h}{2} \{F(b) + F(a)\} + \frac{h^2}{12} \{F'(b) - F'(a)\}$$

$$+ \text{ terms involving } h^4 \quad . \quad . \quad . \quad . \quad . \quad . \quad (135)$$
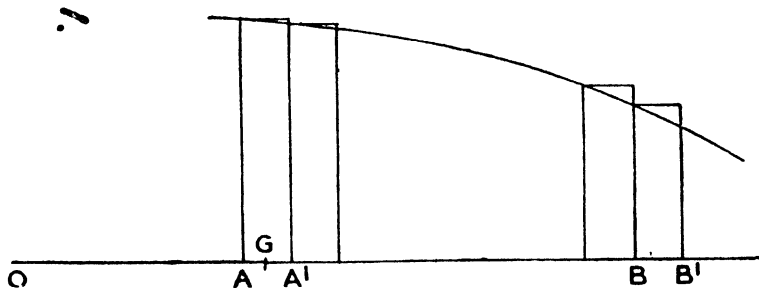
In the figure let OA represent a, OB b, and AA' and BB' h.

AB $= mh$.

$hF(a)$, $hF(b)$ are the rectangular areas on AA', BB'. $h \sum_{a}^{d} F(x)$ is the sum of the rectangular areas on AB.

$\int_{a}^{b} F(x) . dx$ is the curvilinear area on AB, and the term $\frac{h}{2}[F(a) - F(b)]$ is a first approximation for the defect of the curved from the rectilinear area.

———

Some difficulties arise in applying this theorem to the curve of error.

Here $F(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$, and, when x is not great compared with σ, should be represented by a finite vertical length.

$\frac{1}{\sigma} = \frac{1}{\sqrt{pqn}}$, and ∴ $\frac{1}{\sqrt{n}}$, should be finite vertically.

The horizontal distance $AB = mh$ should be finite. It is found in the analysis on pp. 265-8 that the number of successes above or below pn must be considered as of order $\sqrt{n}$; hence m is of order $\sqrt{n}$ and h (the unit step) is of order $\dfrac{1}{\sqrt{n}}$. In other words, in the drawing the rectangles must be supposed so thin that it takes a number of them comparable with $\sqrt{n}$ to give a finite breadth.

In the equation (135) then h is of order $\dfrac{1}{\sqrt{n}}$, $\sum_a^b F(x)$ contains $\sqrt{n}$ terms each finite, and therefore $h\sum_a^b F(x)$ is of order $F(x)$, as is $\displaystyle\int_a^b F(x) \cdot dx$.

The following terms on the right-hand side are successively of orders $\dfrac{1}{\sqrt{n}}$, $\dfrac{1}{n}$ etc. ($F'(b)$ is of course a simple numerical ratio.)

Now give h its value unity, and we have, for the normal curve of error in which terms of order $\dfrac{1}{\sqrt{n}}$ are neglected,

aggregate chance of successes from

$$pn + x_1 \text{ to } pn + x_2 = \int_{x_1}^{x_2} F(x) \cdot dx \quad . \quad . \quad . \quad (136)$$

In the next approximation

$$P_s = P_0 \cdot e^{-\frac{x^2}{2\sigma^2}} \cdot \left\{ 1 - \frac{\kappa}{2}\left(\frac{x}{\sigma} - \frac{x^3}{3\sigma^3}\right)\right\},$$

where terms in $\dfrac{1}{\sqrt{n}}$ are retained and terms in $\dfrac{1}{n}$ neglected. The h term in the formula (135) must be retained.

The result is most conveniently given as the sum of the chances of successes from pn to $pn + x$; for this purpose suppose A in the figure to be a half-unit to left of G, where $OG = pn$ and G is the abscissa of the centre of gravity of the curve (p. 437). Then let $GB = x$.

Sum of chances from G to B $=$ sum from A to B $- \frac{1}{2} \cdot P_0$

$$= \int_0^x P_s \cdot dx + \tfrac{1}{2}(P_x + P_0) - \tfrac{1}{2}P_0.$$

Write $x = z\sigma$.

Hence the sum of chances from

$$0 \text{ to } z\sigma = \frac{1}{\sqrt{2\pi}}\int_0^z e^{-\frac{1}{2}z^2}\left[1 - \frac{\kappa}{2}\left(z - \frac{1}{3}z^3\right)\right]dz + \frac{1}{2\sigma\sqrt{2\pi}}e^{-\frac{1}{2}z^2} \quad (137)$$

when $\frac{1}{\sigma^2}$ is neglected.

Here $\sigma = \sqrt{pqn}, \quad \kappa = \dfrac{q-p}{\sqrt{pqn}}$.

(See Todhunter's *History of the Theory of Probability*, Art. 993.)

5.—*Dr. Sheppard's Corrections for the Moments of Frequency Curves.*

(See *Biometrika*, Vol. III., pp. 308 *seq.*)

Let $y = f(x)$ be the equation of a continuous curve of frequency, whose area is unity.

Let A_p be the area standing on the base $x_p \pm \dfrac{h}{2}$, p being integral, and let the values of A_p for all values of p be known from the observations.

The tth moment computed from the equation of the curve, say $m_t, = \int_a^b x^t . f(x) . dx$, where a and b are the extreme values of x.

The tth moment computed from the observations, when each area is taken as concentrated at the middle point of its grade, say $\mu_t, = \sum_a^b . x_p{}^t . A_p$.

Required to find what correction should be made to μ_t to obtain m_t.

$$A_p = \int_{x_p - \frac{h}{2}}^{x_p + \frac{h}{2}} f(x) . dx = \int_{-\frac{h}{2}}^{\frac{h}{2}} f(x_p + x) . dx$$

$$= \int_{-\frac{h}{2}}^{\frac{h}{2}} \left\{f(x_p) + xf'(x_p) + \frac{x^2}{2!}f^2(x_p) + \ldots\right\} . dx$$

$$= hf(x_p) + \frac{h^3}{24} f^2(x_p) + \frac{h^5}{1920} f^4(x_p) + \ldots.$$

Hence

$$\mu_t = \sum_a^b hx_p{}^t f(x_p) + \sum_a^b \frac{h^3}{24} x_p{}^t f^2(x_p) + \sum_a^b \frac{h^5}{1920} x_p{}^t f^4(x_p) + \ldots$$

$= $ by the Euler-Maclaurin theorem (formula (135))

$$\int_a^b x^t f(x) \cdot dx + \frac{h}{2} \{b^t f(b) + a^t f(a)\} + \frac{h^2}{12} [D(x^t f(x))]_a^b - \frac{h^4}{720} [D^3(x^t f(x))]_a^b$$

$$+ \frac{h^2}{24} \int_a^b x^t f^2(x) \cdot dx + \frac{h^3}{48} \{b^t f^2(b) + a^t f^2(a)\} + \frac{h^4}{288} [D(x^t f^2(x))]_a^b$$

$$+ \frac{h^4}{1920} \int_a^b x^t f^4(x) \cdot dx + \text{terms involving } h^5.$$

Now restrict the investigation to the case where the curve touches the axis at both extremities, so that

$$f(a) = 0 = f(b) = f'(a) = f'(b),$$

and let the contact be so close that also

$$h^2 f^2(a) = h^2 f^2(b) = 0, \text{ and also } h^4 f^3(a) = h^4 f^3(b) = 0,$$

and in all these cases let the presence of a multiplier such as a^t, b^t not make any significant difference from zero.

The expression reduces to

$$\mu_t = \int_a^b x^t f(x) \cdot dx + \frac{h^2}{24} \int_a^b x^t f^2(x) dx + \frac{h^4}{1920} \int_a^b x^t f^4(x) \, dx$$

$$+ \text{terms involving } h^5.$$

Then

$$\int_a^b x^t f^2(x) dx = [x^t f'(x)]_a^b - t \int_a^b x^{t-1} f'(x) \, dx = t(t-1) m_{t-2},$$

and $\int_a^b x^t f^4(x) dx = t(t-1)(t-2)(t-3) m_{t-4}$, by continual integration by parts and use of the conditions.

Since h is generally small in comparison with the moments, terms involving h^5 can be neglected.

$$\therefore \mu_t = m_t + \frac{h^2}{24} t(t-1) m_{t-2} + \frac{h^4}{1920} t(t-1)(t-2)(t-3) m_{t-4}$$

approximately.

Giving t the values 0, 1, 2, 3, 4 in succession, we have

$\mu_0 = m_0 = $ area of curve $= 1$.

$\mu_1 = m_1 = $ zero if the equation is referred to the vertical through the average.

$$\mu_2 = m_2 + \frac{h^2}{12}.$$

$$\mu_3 = m_3 + \frac{h^2}{4} m_1 = m_3 \text{ if } m_1 = 0.$$

$$\mu_4 = m_4 + \frac{h^2}{2} m_2 + \frac{h^4}{80} m_0 = m_4 + \frac{h^2}{2} \left(\mu_2 - \frac{h^2}{12}\right) + \frac{h^4}{80}.$$

$$\therefore \quad m_2 = \mu_2 - \frac{h^2}{12}, \quad \ldots \ldots \ldots \quad (138)$$

$$m_4 = \mu_4 - \frac{h^2}{2}\mu_2 + \frac{7h^4}{240}, \quad \ldots \ldots \quad (139)$$

and $m_1 = \mu_1$, $m_3 = \mu_3$ when the moments are taken about the vertical through the average.

6.—*The Moments and Constants of the Second Approximation to the Generalised Curve of Error.*

The equation to the curve is

$$y = \frac{1}{s\sqrt{2\pi}} e^{-\frac{x^2}{2s^2}} \left\{ 1 - \frac{\kappa}{2}\left(\frac{x}{s} - \frac{1}{3} \cdot \frac{x^3}{s^3}\right) \right\}.$$

Write m_p for $\displaystyle\int_{-\infty}^{\infty} \frac{1}{s\sqrt{2\pi}} e^{-\frac{x^2}{2s^2}} \cdot x^p dx.$

Then $m_0 = 1$, $m_1 = m_3 = \ldots = m_{2p+1} = \ldots = 0$, $m_2 = s^2$,

$$m_{2p} = 1 \cdot 3 \cdot 5 \ldots (2p - 1) \cdot s^{2p} \text{ (formula (23))}.$$

Write M_p for the pth moment of the second approximation. Then

$$M_p = \int_{-\infty}^{\infty} \frac{1}{s\sqrt{2\pi}} e^{-\frac{x^2}{2s^2}} \cdot x^p dx - \frac{\kappa}{2s}\int_{-\infty}^{\infty} \frac{1}{s\sqrt{2\pi}} e^{-\frac{x^2}{2s^2}} \cdot x^{p+1} dx$$

$$+ \frac{\kappa}{6s^3}\int_{-\infty}^{\infty} \frac{1}{s\sqrt{2\pi}} e^{-\frac{x^2}{2s^2}} \cdot x^{p+3} dx$$

$$= m_p - \frac{\kappa}{2s} m_{p+1} + \frac{\kappa}{6s^3} \cdot m_{p+3}.$$

$$\therefore \quad M_{2p} = m_{2p}, \text{ since } m_{2p+1} = 0 = m_{2p+3}$$

and therefore even moments are not affected by the inclusion of the κ term.

$$M_2 = s^2 \quad \ldots \ldots \ldots \quad (140)$$

$$M_{2p+1} = -\frac{\kappa}{2s} m_{2p+2} + \frac{\kappa}{6s^3} \cdot m_{2p+4}, \text{ since } m_{2p+1} = 0$$

$$M_1 = -\frac{\kappa}{2s}\left(m_2 - \frac{1}{3s^2}m_4\right) = -\frac{\kappa}{2s}\left(s^2 - \frac{1}{3s^2} \cdot 3s^4\right) = 0$$

$$M_3 = -\frac{\kappa}{2s}\left(3s^4 - \frac{1}{3s^2} \cdot 15s^6\right) = \kappa \cdot s^3 \quad \ldots \ldots \quad (141)$$

$$M_{2p+1} = -\frac{\kappa}{2s} \cdot 1.3.5 \ldots (2p+1) \cdot s^{2p+2}\left\{1 - \frac{1}{3s^2}(2p+3)s^2\right\}$$

$$= \frac{p}{3} \cdot 1.3.5 \ldots (2p+1) \cdot s^{2p-2} \cdot M_3 \quad \ldots \quad \ldots \quad (142)$$

The origin is the average of the curve, since $M_1 = 0$.

To find the mode we must equate $\frac{dy}{dx}$ to zero.

$$\log\left(ys\sqrt{2\pi}\right) = -\frac{x^2}{2s^2} + \log\left\{1 - \frac{\kappa}{2}\left(\frac{x}{s} - \frac{x^3}{3s^3}\right)\right\} = -\frac{x^2}{2s^2} - \frac{\kappa}{2}\left(\frac{x}{s} - \frac{x^3}{3s^3}\right),$$

since κ^2 is of the order $\frac{1}{n}$ and neglected in the analysis of p. 295.

$$0 = \frac{1}{y} \cdot \frac{dy}{dx} = -\frac{x}{s^2} - \frac{\kappa}{2s}\left(1 - \frac{x^2}{s^2}\right), \quad \ldots \quad \ldots \quad (143)$$

whence $x = -\frac{1}{2}\kappa s$, neglecting κ^2.

$$\therefore \quad \frac{\text{distance that average is to the right of mode}}{s} = \frac{1}{2}\kappa \quad . \quad (144)$$

Then area of the curve standing on the base ON, where ON $= x = zs$, is given by

$$Y_0^z = \frac{1}{\sqrt{2\pi}}\int_0^z e^{-\frac{1}{2}z^2}\left\{1 - \frac{\kappa}{2}\left(z - \frac{z^3}{3}\right)\right\}dz$$

$$= F(z) - \frac{\kappa}{6\sqrt{2\pi}}\{1 - (1 - z^2)e^{-\frac{1}{2}z^2}\}$$

$$= F(z) - \kappa f(z),$$

where $\quad F(z) = \frac{1}{\sqrt{2\pi}}\int_0^z e^{-\frac{1}{2}z^2}dz$

and $\quad f(z) = \frac{1}{6\sqrt{2\pi}}\{1 - (1 - z^2)e^{-\frac{1}{2}z^2}\}.$

These functions are tabulated on p. 271 and p. 303.

$$Y_{-z}^0 = F(z) + \kappa f(z)$$

and the whole chance from $-z$ to $+z$, Y_{-z}^z, is $2F(z)$, as in the normal curve.[*]

---

[*] The corresponding formula from the p, q, n hypothesis, using the Euler-Maclaurin theorem, is $2F(z) + \frac{e^{-\frac{1}{2}z^2}}{s\sqrt{2\pi}}$; but when the data are continuous the last term drops out.

EXAMPLE OF THE SKEW CURVE OF ERROR.

GROUPING OF SCHOOL CHILDREN IN VIᴛʜ GRADE OF *S. LOUIS PUBLIC SCHOOLS ACCORDING TO AGE.

Normal Curve, S = 1.190 (years)
Skew Curve, K = .206
× × × Observations
O .... Centre of Gravity
Mₒ .... Mode
M .... Median
Q_1, Q_2 ...... Quartiles

Scale of Years
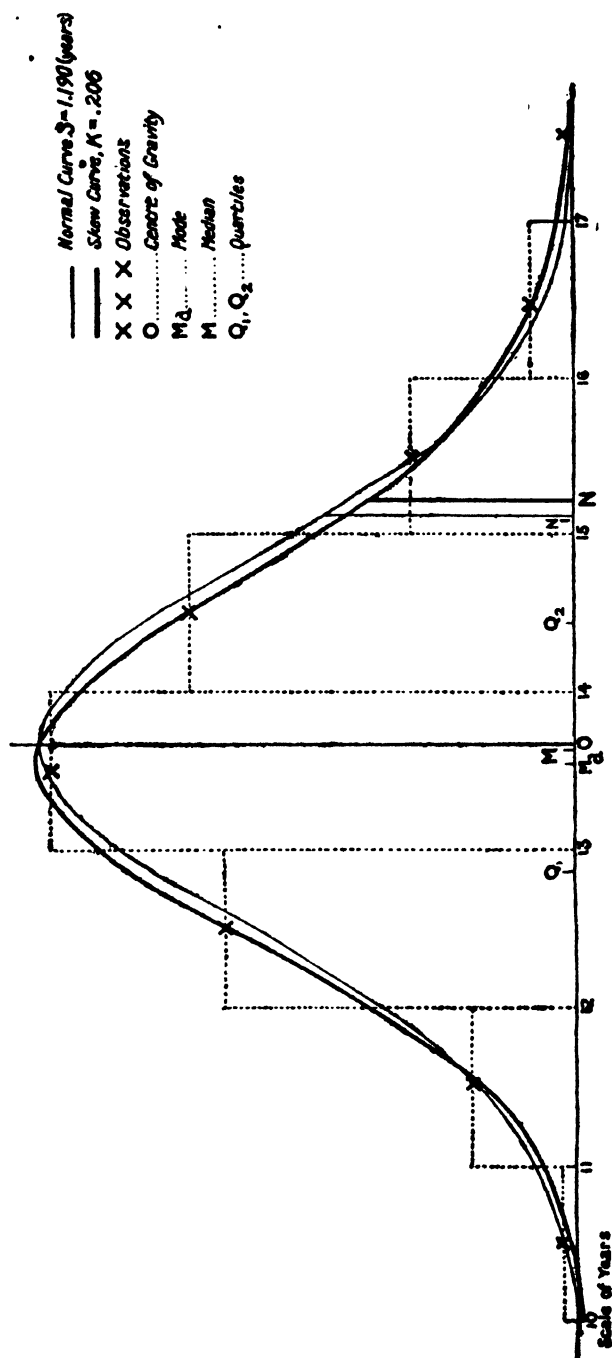
If 'M is the position of the median.we have

$$\tfrac{1}{2} + \text{area on MO} = Y_{-\infty}^{0} \text{, and } \tfrac{1}{2} - \text{area on MO} = Y_{0}^{\infty} \text{,}$$

$$\therefore 2\,\text{area on MO} = Y_{-\infty}^{0} - Y_{0}^{\infty} = \kappa\{f(-\infty)+f(\infty)\} = \frac{\kappa}{3\sqrt{2\pi}}. $$

The ordinates on the small base OM differ from the ordinate at O, viz., $\dfrac{1}{s\sqrt{2\pi}}$, only by terms involving κ.

$$\therefore 2\text{MO} \times \frac{1}{s\sqrt{2\pi}} = \frac{\kappa}{3\sqrt{2\pi}}, \text{ when } \kappa^2 \text{ is neglected,}$$

and $\text{MO} = \tfrac{1}{6}\kappa s = \tfrac{1}{3}$ (distance from mode to average) . . . . (145)

Let the area on MN, where M is the median and N any point, equal the area of the normal curve on ON_1, *i.e.*, $F\left(\dfrac{x_1}{s}\right)$ where $x_1 = ON_1$, and let $NN_1 = v$, where v, as can be seen in the following analysis, is small and of the order κ.

$$ON = x_1 - v.$$

Then

$$F\left(\frac{x_1}{s}\right) = \text{area on MO} + \text{area on } (x_1 - v)$$

$$= \frac{\kappa}{6\sqrt{2\pi}} + \frac{1}{s\sqrt{2\pi}}\int_0^{x_1-v} e^{-\frac{x^2}{2s^2}}\,dx$$

$$- \frac{\kappa}{6\sqrt{2\pi}}\left\{1 - \left(1 - \frac{(x_1-v)^2}{s^2}\right)e^{-\frac{(x_1-v)^2}{2s^2}}\right\}$$

$$= \frac{\kappa}{6\sqrt{2\pi}} + F\left(\frac{x_1}{s}\right) - \frac{v}{s\sqrt{2\pi}}e^{-\frac{x_1^2}{2s^2}} - \frac{\kappa}{6\sqrt{2\pi}} + \frac{\kappa}{6\sqrt{2\pi}}\left(1 - \frac{x_1^2}{s^2}\right)e^{-\frac{x_1^2}{2s^2}},$$

where terms of order v^2, and $v\kappa$ are neglected.

$$\therefore \qquad\qquad v = \frac{\kappa s}{6}\left(1 - \frac{x_1^2}{s^2}\right).$$

The average, s, and κ can be obtained if we know the relative number of observations from the lowest to each of three positions on the horizontal scale, and if we can assume the equation of the frequency curve is that here in question.

The method is most readily explained if we take a numerical example. On p. 309 we have '

Limits of Age.	Number of Children.
0 to 13 years	·296 of 3044
0 to 15 ,,	·867 ,, ,,
0 to 16 ,,	·969 ,, ,,

Let m be the median age, s years the standard deviation, and κs^3 = third moment, all unknown.

In the figure above let M represent the median age and N the age 15 years.

The area on MN is then $·867 - ·500 = F(1·112)$, (p. 271).

Hence

$$ON_1 = 1·112 = \frac{x_1}{s} = z_1$$

$$15 - m = MN = MO + ON_1 - NN_1 = \tfrac{1}{6}\kappa s + x_1 - \tfrac{1}{6}\kappa s\left(1 - \frac{x_1^2}{s^2}\right)$$

$$15 - m = z_1 s + \tfrac{1}{6}\kappa s z_1^2 \text{ where } z_1 = 1·112.$$

Similarly

$$16 - m = z_2 s + \tfrac{1}{6}\kappa s z_2^2, \text{ where } z_2 = 1·866,$$

and

$$m - 13 = z_3 s - \tfrac{1}{6}\kappa s z_3^2, \text{ where } F(z_3) = ·204 \text{ and } z_3 = ·536.$$

A little consideration will show that the negative sign must be taken when N is to the left of M.

We have now three equations for determining m, s, and κ.

$$\tfrac{1}{6}\kappa s (z_1 - z_3) + s = \frac{2}{z_1 + z_3}$$

$$\tfrac{1}{6}\kappa s (z_2 - z_3) + s = \frac{3}{z_2 + z_3}$$

$$\therefore \quad \kappa s = ·278 \quad s = 1·187 \quad \kappa = ·234 \quad m = 13·623$$

$$\text{Average} = m + \tfrac{1}{6}\kappa s = 13·669.$$

(Compare *Statistical Journal*, 1902, pp. 339 to 348.)

From moments depending on the whole nine grades, it was found that $s = 1·190$, $\kappa = ·206$, and average $= 13·665$.

If the average or median is known, or if the curve is known to be normal and $\kappa = 0$, two observations are sufficient for determining the remaining quantities.

Notice that the curves representing the first and second approximations intersect at $x = s\sqrt{3}$.

The area of the skew curve standing on MN = the area of the normal curve standing on ON when $x = \pm\, s$.

The excess of the skew curve over the normal curve for any distance ON to the right = the defect for the same distance to the left.

7.—Ratio of Unweighted Averages.

Let $M_1, M_2 \ldots M_n$ be the true measurements of n quantities at one time, and $M_1', M_2' \ldots$ of similar quantities at another time.

Let $n\overline{m} = SM_t$, $M_t = \overline{m} + m_t$, $Sm_t = 0$, $n\overline{m}' = SM_t'$, $M_t' = \overline{m}' + m_t'$, $Sm_t' = 0$, $n\sigma_m{}^2 = Sm_t{}^2$, $n\sigma_{m'}{}^2 = Sm'_t{}^2$.

Let $\overline{m}' = \overline{m}\,(1 + \rho)$, $M_t' = (1 + a + u_t)M_t$ where $Su_t = 0$, and

$$\therefore \qquad\qquad \rho = a + \frac{Sm_t u_t}{n\overline{m}}.$$

Here $\bar{u}$ measures the mean of the ratio increases of the quantities, and ρ measures the ratio increase in the mean of the quantities. These tend to be equal, if the larger quantities are not on the whole subject to the larger increases, or conversely.

Suppose the quantities to be erroneously measured as $M_t(1 + e_t)$ and $M_t'(1 + e_t')$ etc. Then by formula (70) the standard

deviations of the errors in m and m' are $\dfrac{\sigma}{\sqrt{n}}\sqrt{\left(1 + \dfrac{\sigma_m{}^2}{\overline{m}^2}\right)}$ and

$\dfrac{\sigma_1}{\sqrt{n}}\sqrt{\left(1 + \dfrac{\sigma_{m'}{}^2}{\overline{m}'^2}\right)}$, where σ and σ_1 are typical of e_t and e_t'.

If the errors in the two sets of measurements are independent of each other, then (by p. 318, formula (63)),

$$s_r{}^2 = \frac{1}{n}\left\{\sigma^2\left(1 + \frac{\sigma_m{}^2}{\overline{m}^2}\right) + \sigma_1{}^2\left(1 + \frac{\sigma_{m'}{}^2}{\overline{m}'^2}\right)\right\},$$

where s_r is the standard deviation of errors in $\dfrac{\overline{m}'}{\overline{m}}$, i.e. in $1 + \rho$.

It is frequently the case, however, that the error e_t' in the measurement of M'_t is of the same sign and not far from the same magnitude as e_t, the error in the earlier measurement of the corresponding M_t.

Write $\qquad\qquad d_t = e_t' - e_t.$

Then if e is the resulting error in the ratio of the averages

$$\frac{\overline{m}'}{\overline{m}}(1 + e) = \frac{S\{M_t'(1 + e_t')\}}{S\{M_t(1 + e_t)\}}$$

$$e = \frac{S\{M_t'(1 + e_t')\}.SM_t - S\{M_t(1 + e_t)\}.SM_t'}{S\{M_t(1 + e_t)\}.SM_t'}$$

$$= \frac{\overline{m}S(M_t'e_t') - \overline{m}'.S(M_te_t)}{\overline{m}'S\{M_t(1 + e_t)\}}$$

$$= \frac{\overline{m}S(M_t'd_t) + S\{(\overline{m}M_t' - \overline{m}'M_t)e_t\}}{n\overline{m}\overline{m}'},$$

neglecting e_t^2 and e_te_t',

$$= \frac{S(M_t'd_t)}{n\overline{m}'} + \frac{S\{(u_t + a - \rho)M_te_t\}}{n\overline{m}'} = \frac{S(M_t'd_t)}{n\overline{m}'} + \frac{SM_tu_te_t}{n\overline{m}'},$$

if $\overline{u} - \rho$ is neglected.

Hence if s_r is the standard deviation of e, and σ_d, σ the standard deviations of d_t and e_t, or their weighted standard deviations if they are not all from identical frequency curves,

$$s_r^2 = \frac{1}{n^2\overline{m}'^2}\sigma_d^2 . S(M_t'^2) + \frac{1}{n^2\overline{m}'^2}.\sigma^2.S(M_t^2u_t^2), \text{ by formula (55)},$$

if d_t and e_t are uncorrelated.

Now $S(M_t')^2 = S(\overline{m}' + m_t')^2 = n(\overline{m}'^2 + \sigma_m^2)$,

and $S(M_t^2u_t^2) = n\overline{m}^2\sigma_u^2 + n\sigma_m^2\sigma_u^2 + Su_t^2(m_t^2 - \sigma_m^2) + 2\overline{m}S(m_tu_t^2)$,

where $n\sigma_u^2 = Su_t^2$,

$$= n\sigma_u^2(\overline{m}^2 + \sigma_m^2) + \text{ terms which tend to be negligible.}$$

$$\therefore \quad s_r^2 = \frac{1}{n}\sigma_d^2\left(1 + \frac{\sigma_m'^2}{\overline{m}'^2}\right) + \frac{1}{n}.\sigma^2.\left(1 + \frac{\sigma_m^2}{\overline{m}^2}\right).\sigma_u^2.\frac{1}{(1 + \rho)^2}$$

approx. . . . . . . (146)

If e_t and e_t' were independent σ_d^2 would equal $\sigma^2 + \sigma_1^2$, while if $e_t = e_t'$ etc. σ_d would be zero. Hence σ_d may be regarded as between 0 and $\sigma\sqrt{2}$.

The magnitude of the second term depends on σ_u, which measures the variation in the rates of increase of the different quantities, and is known from the observations.

Hence if similar errors are made in observations at both dates of quantities which increase at nearly the same rates, the error in the ratio of the computed averages is small, and, if n is great, very small.

8.—*Ratio of Weighted Averages.*

In the case of weighted averages the formula becomes more complex.

Let $W_t = \bar{w} + w_t$, and $W_t' = \bar{w}' + w_t'$ be any pair of weights at the two dates, where $\bar{w}$, $\bar{w}'$ are the averages of the weights. Write $n\sigma_w^2 = Sw_t^2$ and $n\sigma_{w'}^2 = Sw_t'^2$.

Let $W_t' = W_t(1 + \bar{v} + v_t)$, where $Sv_t = 0$, and write $n\sigma_v^2 = Sv_t^2$.

Suppose $W_t(1 + \eta_t)$ and $W_t'(1 + \eta_t')$ to be taken in error for W_t, W_t', and write σ' for the standard deviation of η_t.

Let $$\bar{m}_w = \frac{S(W_t M_t)}{SW_t} \text{ and } \bar{m}_w' = \frac{S(W_t' M_t')}{SW_t'}.$$

Other letters have the same meaning as in the previous note.

Required the error in $\dfrac{\bar{m}_w'}{\bar{m}_w}$, say e.

$$\frac{\bar{m}_w'}{\bar{m}_w}(1 + e) = \frac{S\{W_t'(1 + \eta_t')M_t'(1 + e_t')\}}{S\{W_t(1 + \eta_t)M_t(1 + e_t)\}} \cdot \frac{S\{W_t(1 + \eta_t)\}}{S\{W_t'(1 + \eta_t')\}},$$

and hence, after reduction in which products and squares of η_t, e_t are neglected,

$$e = \frac{S(W_t' M_t' e_t')}{n\bar{w}'\bar{m}_w'} - \frac{S(W_t M_t e_t)}{n\bar{w}\bar{m}_w} + \frac{S\{W_t(\bar{m}_w - M_t)\eta_t\}}{n\bar{w}\bar{m}_w} - \frac{S\{W_t'(\bar{m}_w' - M_t')\eta_t'\}}{n\bar{w}'\bar{m}'_w}$$

$$\dots \dots \dots \dots \dots (147)$$

To obtain approximate results neglect all sums of products, where the sum of the factors of one kind is zero. This leads to taking $\bar{m}_w = \bar{m}$, $\bar{m}_w' = \bar{m}'$, $\bar{u} = \rho$, $\bar{w}' = (1 + \bar{v})\bar{w}$, and to further simplifications in the reduction.

Write $d_t' = \eta_t' - \eta_t$ and σ_d' for its standard deviation.

Then $$e = \frac{S(W_t' M_t' d_t)}{n\bar{w}'\bar{m}'} + \frac{S(W_t' M_t u_t)}{n\bar{w}'\bar{m}'}e_t + \frac{S(W_t M_t v_t)}{n\bar{w}'\bar{m}}e_t$$
$$+ \frac{S(W_t' m_t' d_t')}{n\bar{w}'\bar{m}} + \frac{S(W_t' M_t u_t)}{n\bar{w}'\bar{m}'}\eta_t + \frac{S(W_t m_t v_t)}{n\bar{w}'\bar{m}}\eta_t.$$

Hence approximately

$$s_r^2 = \frac{1}{n}\left(1 + \frac{\sigma_w^2}{\bar{w}'^2}\right)\left(1 + \frac{\sigma_m'^2}{\bar{m}'^2}\right)\sigma_d^2 + \frac{1}{n}\left(1 + \frac{\sigma_w'^2}{\bar{w}'^2}\right)\left(\frac{\sigma_m'^2}{\bar{m}'^2}\right)\sigma_d'^2$$
$$+ \frac{1}{n}\left(\frac{\sigma_u}{1 + \bar{u}}\right)^2\left(1 + \frac{\sigma_w'^2}{w'^2}\right)\left(1 + \frac{\sigma_m^2}{\bar{m}^2}\right)\left(\sigma^2 + \sigma'^2\right)$$
$$+ \frac{1}{n}\left(\frac{\sigma_v}{1 + \bar{v}}\right)^2\left(1 + \frac{\sigma_w^2}{\bar{w}^2}\right)\left\{\left(1 + \frac{\sigma_{1n}^2}{\bar{m}^2}\right)\sigma^2 + \frac{\sigma_{1n}^2}{\bar{m}^2}\sigma'^2\right\} \quad (148)$$

The terms involving σ^2, $\sigma_d{}^2$ (which measure the errors in the quantities M_t) are similar to those when the average is unweighted, except for a factor (greater than 1 and generally less than 2) involving weights, and a term involving also the small factor $\sigma_\bullet{}^2$ which measures the variation in the change of weights.

Of the three terms involving σ'^2, $\sigma_{d'}{}^2$ (which measure the errors in weights) the first and the third contain the factors $\left(\dfrac{\sigma_{m'}}{\bar{m}'}\right)^2$ and $\left(\dfrac{\sigma_m}{\bar{m}}\right)^2$ respectively, which are small when the M's are little dispersed, and the second involves $\left(\dfrac{\sigma_u}{1 + \bar{u}}\right)^2$ which is small when the rates of increase of the quantities are nearly equal.

The actual values of all the coefficients of σ, σ', σ_d, σ_d' can be obtained from the observations, and their relative importance discovered ; but we can say without evaluation that when quantities little dispersed increase at rates not far from equal, errors in weights have little importance as compared with equal errors in quantities.

In such cases a first approximation would be

$$s_r = \frac{\sigma_d}{\sqrt{n}}\sqrt{\left(1 + \left(\frac{\sigma_w}{\bar{w}}\right)^2\right)}, \quad \cdots \quad (149)$$

but if $\dfrac{\sigma_u}{1 + \bar{u}}$ is not small, a better approximation would be

$$s_r = \frac{1}{\sqrt{n}} \cdot \sqrt{\left[\left(1 + \left(\frac{\sigma_w}{\bar{w}}\right)^2\right)\left\{\sigma_d{}^2 + \left(\frac{\sigma_u}{1 + \bar{u}}\right)^2\left(\sigma^2 + \sigma'^2\right)\right\}\right]} \quad \cdot \quad (150)$$

It is seldom that σ_d, which measures the difference of errors, is small compared with one error, though it is likely to be less than $\sqrt{2}.\sigma$.

It is advisable to test the coefficients roughly from the observations before neglecting terms ; and also where there are any signs that the neglected products are not small, or any of the errors are likely to be specially large, the unabridged form (147) should be used.

(See *Statistical Journal*, 1911–12, pp. 81–88, " Measurement of the Accuracy of an Average.")

9.—*Normality of Standard Deviations of the Errors in Moments, etc.*

[Based on Sheppard's "Application of the Theory of Error," *Transactions* of the Royal Society, Vol. 192, 1898, A. 229, pp. 117–128, but with modifications in notation and treatment.]

In a universe containing N things p_1N are at x_1, p_2N at $x_2 \ldots$ $p_1 + p_2 + \ldots = 1$, $F = a_1 p_1 + a_2 p_2 + \ldots$ where a_1, $a_2 \ldots$ are constants.

In a selection of n things, n_1 are found at x_1, n_2 at $x_2 \ldots$, $n_1 + n_2 + \ldots = n$.

Write $F + f = a_1 \dfrac{n_1}{n} + a_2 \dfrac{n_2}{n} + \ldots$

$$f = a_1 \frac{n_1}{n} + a_2 \frac{n_2}{n} + - F\left(\frac{n_1}{n} + \frac{n_2}{n} + \ldots\right) = b_1 \frac{n_1}{n} + b_2 \frac{n_2}{n} + \ldots,$$

where $b_1 = a_1 - F$ etc.

Then $Sb_\iota p_\iota = Sa_\iota p_\iota - F \cdot Sp_\iota = F - F = 0$

$Sb_\iota^2 p_\iota = Sa_\iota^2 p_\iota - 2F Sa_\iota p_\iota + F^2 \cdot Sp_\iota = Sa_\iota^2 p_\iota - F^2.$

Required to find M_s, $=$ mean f^s, and to show that its relation to M_2, $=$ mean f^2, is that found in the normal curve of error.

The expression

$$E = \left(p_1 e^{b_1 \frac{s}{n}} + p_2 e^{b_2 \frac{s}{n}} + \ldots\right)^n,$$

expanded by the multinomial theorem, gives the sum of terms

$$\frac{n!}{n_1! \, n_2! \ldots}\left(p_1 e^{b_1 \frac{s}{n}}\right)^{n_1}\left(p_2 e^{b_2 \frac{s}{n}}\right)^{n_2} \ldots,$$

subject to the condition $n_1 + n_2 + \ldots = n$

$\qquad\qquad = $ sum of terms $P \cdot e^{fs}$,

where $\qquad\qquad f = b_1 \dfrac{n_1}{n} + b_2 \dfrac{n_2}{n} + \ldots,$

and $\qquad\qquad P = \dfrac{n!}{n_1! \, n_2! \ldots} p_1{}^{n_1} p_2{}^{n_2} \ldots$

and is the whole chance that the selection n_1 at x_1, n_2 at $x_2 \ldots$ should be made, as may be seen by expanding the multinomial

$$(p_1 + p_2 + \ldots)^n$$

$$\therefore E = \text{sum of } P\left(1 + af + \frac{a^2}{2}f^2 + \ldots + \frac{a^s}{s\,|}f^s + \ldots\right)$$

$$= M_0 + aM_1 + \frac{a^2}{2}M_2 + \ldots + \frac{a^s}{s\,|}M_s + \ldots$$

Also $E = \left(S p_t + \frac{a}{n}Sb_t p_t + \frac{a^2}{2n^2}Sb_t^2 p_t + \frac{a^3}{6n^3}C_3 + \frac{a^4}{24n^4}C_4 + \ldots\right)^n$

from the expansion of the terms $e^{b_t \frac{a}{n}} \ldots$, where

$$C_3 = Sb_t^3 p_t, \; C_4 = Sb_t^4 p_t \ldots, \text{ and } Sp_t = 1, \; Sb_t p_t = 0.$$

$$\therefore \qquad E = (1 + \frac{a^2}{2n^2}Sb_t^2 p_t + \ldots)^n.$$

Equating the first three coefficients in the two expressions for E.

$$M_0 = 1, \; M_1 = 0, \; M_2 = \frac{1}{n}Sb_t^2 p_t.$$

$$\therefore 1 + \frac{a^2}{2}M_2 + \ldots + \frac{a^s}{s\,|}M_s + \ldots = \left(1 + \frac{a^2}{2n}M_2 + \frac{a^3}{6n^3}C_3 + \ldots\right)^n.$$

Now when n is large, and $M_2, \dfrac{C_3}{n^{\frac{1}{2}}}, \dfrac{C_4}{n^2}, \dfrac{C_5}{n^{\frac{1}{2}}} \ldots$ are finite, we have,

if we neglect $\dfrac{1}{\sqrt{n}}$,

$$1 + \frac{a^2}{2}M_2 + \ldots + \frac{a^s}{s\,|}M_s + \ldots = \left(1 + \frac{a^2}{2n}M_2\right)^n = e^{\frac{a^2}{2}M_2}$$

$$= 1 + \frac{a^2}{2}M_2 + \ldots + \frac{a^{2t}}{t\,|\,2^t}\left(M_2\right)^t + \ldots$$

Hence, in this case, $M_s = 0$ if s is odd, and $M_{2t} = \dfrac{(2t)\,!}{t\,|\,2^t}\left(M_2\right)^t$, as in the normal curve of error.

The conditions that $\dfrac{C_3}{n^{\frac{1}{2}}}$ etc. are finite are similar to those in the Edgeworthian analysis on pp. 295 *seq.*, but need consideration for each case to which the theorem is applied.

Thus on pp. 419-20 $f = x_1^2 e_1 + x_2^2 e_2 + \ldots$, where e_1 is $\dfrac{n_1}{n} - p_1$, $F = \mu_2$ the second moment of the universe from which the selection is made, and $b_t = x_t^2 - \mu_2$.

$$M_2 = \frac{1}{n}S(x_t^2 - \mu_2)^2 p_t = \frac{1}{n}(\mu_4 - \mu_2^2)$$

$$C_3 = S(x_t^2 - \mu_2)^3 p_t = (\mu_6 - 3\mu_4\mu_2 + 2\mu_2^3)$$

$$\frac{C_3}{n^{\frac{1}{2}}} = M_2^{\frac{3}{2}} \cdot \frac{\mu_6 - 3\mu_4\mu_2 + 2\mu_2^3}{(\mu_4 - \mu_2^2)^{\frac{3}{2}}}.$$

Similarly

$$\frac{C_4}{n^2} = \left(M_2\right)^2 \frac{\mu_8 - 4\mu_6\mu_2 + 6\mu_4\mu_2{}^2 - 3\mu_2{}^4}{(\mu_4 - \mu_2{}^2)^2} \text{ etc.}$$

Now if the ratios $\frac{\mu_4}{\sigma^4}, \frac{\mu_6}{\sigma^6}, \frac{\mu_8}{\sigma^8} \ldots$ are finite, where $\sigma^2 = \mu_2$, we have that $\frac{C_2}{n^{\frac{3}{2}}}, \frac{C_4}{n^2} \ldots$ are finite, as was required.

Hence if the curve of frequency of the universe satisfies these conditions, which correspond in fact to a reasonable concentration about the average, with no groups of importance beyond a small multiple of σ, the curve of frequency for the errors of the second moment (and of the standard deviation) are normal.

A similar but simpler analysis shows the errors of the average have normal frequency ($f = x_1e_1$ etc. $F = 0$, $b_t = x_t$).

In the case of the analysis of the correlation coefficient (p. 422)

$$b_t = r\left(\frac{x_ty_t}{M} - \frac{x_t{}^2}{2\lambda} - \frac{y_t{}^2}{2\mu} - 1 + \tfrac{1}{2} + \tfrac{1}{2}\right), \text{ and } p_t = z_t.$$

$$M_2 = \frac{1}{n} S r^2 \left(\frac{x_ty_t}{M} - \frac{x_t{}^2}{2\lambda} - \frac{y_t{}^2}{2\mu}\right)^2 z_t$$

$$= \frac{r^2}{n}\left(\frac{M_{22}}{M^2} + \frac{\lambda_4}{4\lambda^2} + \frac{\mu_4}{4\mu^2} - \frac{M_{31}}{M\lambda} - \frac{M_{13}}{M\mu} + \frac{M_{22}}{2\lambda\mu}\right)$$

$$C_3 = S r^3 \left(\frac{x_ty_t}{M} - \frac{x_t{}^2}{2\lambda} - \frac{y_t{}^2}{2\mu}\right)^3 z_t = r^3 \cdot \left(\frac{M_{33}}{M^3} + \frac{\lambda_6}{8\lambda^3} + \ldots\right).$$

Writing $\lambda = \sigma_1{}^2$, $\mu = \sigma_2{}^2$, we have, if $\frac{\lambda_s}{\sigma_1{}^s}, \frac{\mu_t}{\sigma_2{}^t}, \frac{M_{st}}{\sigma_1{}^s\sigma_1{}^t}$ are finite for all values of s and t, then $\frac{C_3}{n^{\frac{3}{2}}} = M_2{}^{\frac{3}{2}} \times$ finite quantity, and higher terms can be similarly dealt with as before.

Hence if the moments and products of the two-dimensional frequency distributions satisfy the conditions already described, the error curve of the correlation coefficient is normal.

10.—*The Method of Least Squares.*

This is a method that has for a long time been used for assigning the values to be taken when there are a number of inexact measurements at choice.

Suppose a quantity s to be related to k unknown constants $x_1, x_2 \ldots x_k$ by the equation $s = u_1x_1 + u_2x_2 + \ldots + u_kx_k$, where

u_1, u_2 ... are quantities that can be observed; and let n sets of observations be made giving

$$_1z = {}_1u_1x_1 + {}_1u_2x_2 + \ldots + {}_1u_kx_k$$
$$\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot$$
$$_nz = {}_nu_1x_1 + {}_nu_2x_2 + \ldots + {}_nu_kx_k$$

where the z's and u's are known.

If $n = k$, the x's can be exactly determined. If $n < k$, there are an infinite number of solutions and the equations are indeterminate.

If $n > k$ the equations are in general inconsistent, and the problem is to assign values to x_1, x_2 ... which minimise the inconsistency, which is assumed to be due to imperfect measurements of the u's.

Write d_1, d_2, ... for the differences between $_1z$, $_2z$ ... and the values obtained from true values of x_1, x_2 ..., say X_1, X_2 ...

Then
$$_1u_1X_1 + {}_1u_2X_2 + \ldots + {}_1u_kX_k - {}_1z = d_1$$
$$_2u_1X_1 + {}_2u_2X_2 + \ldots + {}_2u_kX_k - {}_2z = d_2$$
$$\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot$$

It is assumed that d_1, d_2 ... are errors whose chances are given by a normal curve $P = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{d^2}{2\sigma^2}}$. The assumption is generally based on demonstrations that under certain hypotheses as to the nature of accidental errors this normal form is obtained. Whatever may be the validity of these hypotheses in physical or geodetical measurements, it is not safe to assume that they apply to statistical or biometric measurements, whether of deviations from an average or errors due to sampling.

The solution is obtained by finding those values of X_1, X_2 ... which make the probability that d_1, d_2 ... would occur together a maximum, that is which make the sum of $d_1^2 + d_2^2 + \ldots$ à minimum. Write $f(d_1, d_2 \ldots)$ for this sum.

The conditions for a minimum are $\dfrac{\partial f}{\partial X_1} = 0 = \dfrac{\partial f}{\partial X_2} = \ldots$

These give

$$\frac{1}{2} \cdot \frac{\partial f}{\partial X_1} = \underset{t=1}{\overset{t=n}{S}} {}_tu_1 ({}_tu_1X_1 + {}_tu_2X_2 + \ldots - {}_tz) = 0$$

$$\frac{1}{2} \cdot \frac{\partial f}{\partial X_2} = \underset{t=1}{\overset{t=n}{S}} {}_tu_2 ({}_tu_1X_1 + {}_tu_2X_2 + \ldots - {}_tz) = 0$$

$$\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot$$
$$\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot$$

which may be written

$$X_1 . Su_1{}^2 + X_2 . Su_1 u_2 + \ldots = Su_1 z$$
$$X_1 Su_1 u_2 + X_2 . Su_2{}^2 + \ldots = Su_2 z$$
$$\cdots \cdots \cdots \cdots \cdots$$
$$X_1 Su_1 u_k + X_2 Su_2 u_k + \ldots = Su_k z,$$

k equations giving the k quantities X_1, $X_2 \ldots$

This method is found in practice to give quite generally good empirical values of the unknown quantities. It is used above on pp. 239, 240 in its simple form, and a corresponding method where the validity can be tested is used on pp. 364-5.

(See Merriman, *Method of Least Squares*, and Weld, *Theory of Errors and Least Squares*.)

  11.—*Simpler Method of Obtaining Formula* (130), *p.* 429.

  Beginning at the top of p. 428 we may write by the Multinomial Theorem that the joint chance of finding the errors $e_1 \ldots e_i \ldots e_n$ is

$$P = \frac{N !}{(m_1 + e_1) ! \ldots (m_i + e_i) ! \ldots (m_n + e_n) !} \cdot$$
$$p_1{}^{m_1 + e_1} \ldots p_i{}^{m_i + e_i} \ldots p_n{}^{m_n + e_n}.$$

Suppose the greatest of the terms $m_i{}^{-1}$ to be negligible, and apply Stirling's formula to the factorials. After reduction we have

$$P = K . \Pi \left(1 + \frac{e_i}{m_i} \right)^{- m_i - e_i - \frac{1}{2}},$$

where $K^{-1} = (2\pi)^{\frac{n-1}{2}} (m_1 \ldots m_i \ldots m_n)^{\frac{1}{2}} N^{-\frac{1}{2}}$.

  Then $\log (P \div K) = - \Sigma (m_i + e_i + \frac{1}{2}) \log \left(1 + \frac{e_i}{m_i} \right)$

$$= - \Sigma (m_i + e_i + \frac{1}{2}) \left(\frac{e_i}{m_i} - \frac{1}{2} \frac{e_i{}^2}{m_i{}^2} + \ldots \right)$$

$$= - \Sigma e_i - \frac{1}{2} \Sigma \frac{e_i{}^2}{m_i} \pm \text{ terms of order } m_i{}^{-\frac{1}{2}},$$

since e_i is of order $m_i{}^{\frac{1}{2}}$.

  If terms of order $m_i{}^{-\frac{1}{2}}$ are negligible, then since $\Sigma e_i = 0$,

$$P = K e^{-\frac{1}{2} \Sigma \frac{e_i{}^2}{m_i}},$$

as in formula (130).

SUPPLEMENTS.

SUPPLEMENT I. KURTOSIS AS MEASURED BY κ_2. ILLUSTRATIONS. (See p. 252.)

κ_2 cannot be less than 1, since $n\ (a^4 + b^4 + \ldots$ to n terms) is greater than $(a^2 + b^2 + \ldots$ to n terms)^2, unless $a = b = \ldots$ when it equals 1. There is no upper limit to it.

Eight diagrams (A) are drawn, so that all their areas are equal and all their standard deviations equal, each diagram being symmetrical, which yield ascending values of κ_2 from $1 \cdot 1$ to $9 \cdot 0$.

In diagrams 1 and 8, out of $m + n$ observations m are at zero and $\frac{1}{2}n$ at unit distance to left and right. In such a case $\kappa_2 = 1 + m/n$. When $m = 0$, $\kappa_2 = 1$; as m increases, κ_2 increases without limit (n decreasing, since $m + n$ is kept constant). In 1, $m = \frac{1}{11}$, $n = \frac{10}{11}$, $\kappa_2 = 1 \cdot 1$. In 8, $m = \frac{8}{9}$, $n = \frac{1}{9}$, $\kappa_2 = 9$.

In diagram 2 the distribution is graded upwards from zero at the centre to the extremes; $\kappa_2 = 1\frac{1}{3}$.

In diagram 3 the observations are uniformly distributed and $\kappa_2 = 1 \cdot 8$.

Diagram 4 is a half ellipse, where the observations are more numerous at the centre than at the extremes, $\kappa_2 = 2$.

In diagram 5 the number of observations increases uniformly from the extremes to the centre. $\kappa_2 = 2 \cdot 4$.

Diagram 6 shows the normal curve of error, with $\kappa_2 = 3$.

In diagram 7 the distribution is two elliptical quadrants placed so as to touch at the central vertical. $\kappa_2 = 3 \cdot 22$. (Exactly $2(992 - 315\,\pi)\ (4 - \pi) \div 15(16 - 5\pi)^2$).

All these results can be verified by direct integration.

SUPPLEMENT II. CORRECTION FOR CRUDE VALUE OF THE MEAN DERIVATION.

(See pp. 111, 253, 439.)

When the observations are graded, as in the example on p. 253, there is some difficulty in computing the mean deviation. The crude method is to add the deviations measured
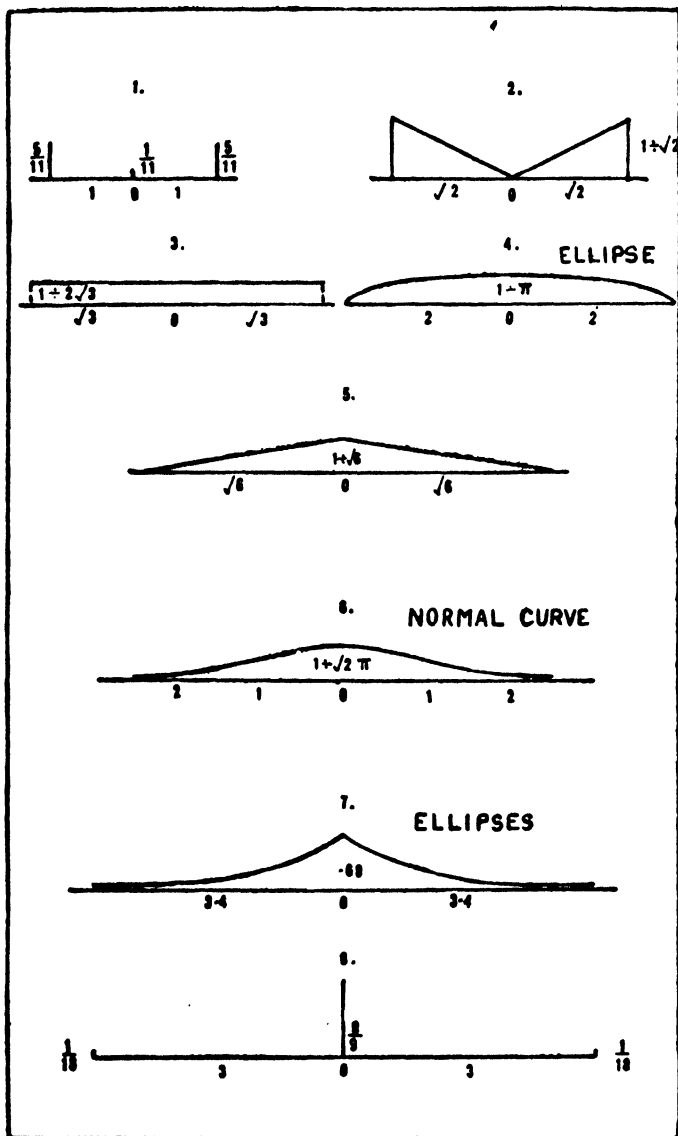
DIAGRAM A.

from the centre of the grade in which the average or median is situated, disregarding their signs, and to divide by the number of observations. Thus we have $(4906 + 4032) \div 3404 = 2 \cdot 6257$ (unit 5 lbs.).

This process requires three corrections, one of which is analogous to Sheppard's, of which the result is valid to the same extent and under the same conditions as apply to the use of his formulæ (pp. 439 *seq.*). The second is to allow for the contribution of the central grade, the third to the distance from the centre of that grade to the point from which deviations are measured.
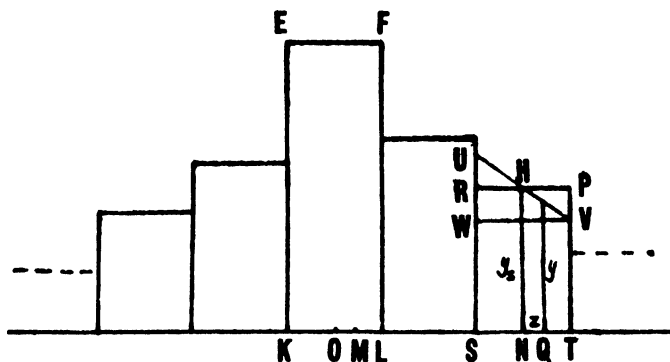
ON $= x_a$, NH $= y_a$, NQ $= z$, y is the vertical from Q to UV.

DIAGRAM B.

1. In any grade, ST, breadth h, the number of observations is (proportional to) SRPT $= h \times y_s$, where y_s is the height SR.

In the crude method of computing the mean deviation these observations are taken as concentrated at N, the middle point of ST, and their contribution to the sum of the deviations about O (mid-point of the central grade KL) is $h y_s x_s$, where ON $= x_s$.

If, in fact, the observations are distributed in a curve $y = f(x)$, as indicated by UV, the true contribution of this group may be obtained as follows.

Take N as a temporary origin, write NQ $= z$, where Q is any point between S and T.

The sum of the deviations arising from the grade ST is

$$\int_{-\frac{h}{2}}^{\frac{h}{2}} y\,(x_s + z)\,dz = \int_{-\frac{h}{2}}^{\frac{h}{2}} \{f\,(x_s) + zf'(x_s) + \ldots\}\,(x_s + z)\,dz$$
$$= hx_s f(x_s) + \tfrac{1}{12} h^3 . f'(x_s) + \ldots$$
$$= hx_s y_s + \tfrac{1}{12} h^2 . hf'(x_s) + \ldots$$
$$= hx_s y_s - \tfrac{1}{12} h^2 UW, \text{ approx.}$$

where VW is horizontal, UV is taken as a straight line (neglecting $f''(x_s)$), and therefore $f'(x_s) = -UW/h$. To this order of approximation UV bisects RP at H.

The contribution of each grade is therefore over-estimated by such an amount as $h^2 UW/12$. The same result is reached in the left of the figure.

Take h as the same for all grades.

If we suppose that the height to right and left diminishes, so that such lines as UV reach the horizontal axis with sufficient approximation, and that the net adjustment due to curvature in the central grade is negligible, then the sum of UV's on each side is approximately FL or EK $= N_0/h$, where N_0 is the number of observations in the central grade.

We have therefore to subtract

$$2 \times \frac{h^2}{12} \times \frac{N_0}{h} = \tfrac{1}{6} h N_0 \quad . \quad . \quad . \quad . \quad \text{(i)}$$

from the sum of the deviations crudely measured.

(2) The sum of the deviations in the central grade about O, ignoring curvature, is

$$2 \times \tfrac{1}{2} N_0 \times \tfrac{1}{4} h = \tfrac{1}{4} h N_0 \quad . \quad . \quad . \quad \text{(ii)},$$

which is to be added to the crude sum. (i) and (ii) together lead to addition of $\tfrac{1}{12} h N_0$.

(3) Now measure the deviations from M between K and L, where OM $= d$. Let N_1 and N_{-1} be the number of observations to the right of L and left of K respectively.

We have to add $N_{-1}d$ and subtract $N_1 d$ for the outside grades.

The whole contribution of the central grade is now

$$N_0/h\left\{\left(\frac{h}{2} + d\right) . \tfrac{1}{2}\left(\frac{h}{2} + d\right) + \left(\frac{h}{2} - d\right) . \tfrac{1}{2}\left(\frac{h}{2} - d\right)\right\} = N_0\left(\frac{h}{4} + \frac{d^2}{h}\right),$$

instead of $\tfrac{1}{4} h N_0$ as in (ii).

In all we have to add $\left(N_{-1} - N_1 + \dfrac{d}{h}N_0\right)d$ . . . (iii)

Assemble these results, writing N for the whole number of observations, η_0 for the crude and η for the corrected mean deviation.

$$N\eta = N\eta_0 + \tfrac{1}{12}hN_0 + \left(N_{-1} - N_1 + \dfrac{d}{h}N_0\right)d$$

This is a general result.

Write η_m for the mean deviation from the median.

Here $\tfrac{1}{2}N = N_{-1} + \left(\dfrac{h}{2} + m\right)N_0/h = N_1 + \left(\dfrac{h}{2} - m\right)N_0/h,$

where m is the distance from O to the median.

$\therefore N_{-1} - N_1 = -2mN_0/h.$

Write $N_0 = lN$, so that l is the proportion of the observations that fall in the central grade.

We have $\eta_m = \eta_0 + l\left(\dfrac{h}{12} - \dfrac{m^2}{h}\right).$

For any other point in the same central grade

$$\eta = \eta_0 + l\left(\dfrac{h}{12} + \dfrac{d^2}{h}\right) - 2dml/h$$
$$= \eta_m + l(d - m)^2/h. . . . . . \text{(iv)}$$

In the case $d = \bar{x}$, the distance the average is to the right of O, this applies to deviations from the average, so that

$$\eta_a = \eta_m + l(\bar{x} - m)^2/h . . . . \text{(v)}$$

In the example on p. 253, the median is approximately $102\tfrac{1}{2} - \tfrac{10}{100}$ (of 5 lbs.) $h = 1$, if we take 5 lbs. as the unit. $m = -\tfrac{10}{101} = -·099.$

$\eta_o = 8938 \div 3404 = 2·6257,\ l = 404 \div 3404 = ·119$

$\eta_m = 2·6257 + ·119\left(\dfrac{1}{12} - \dfrac{10^2}{101^2}\right) = 2·6344$

$\bar{x} = ·2568$

$\eta_a = \eta_m + ·119(·2568 - ·099)^2 = 2·6344 + ·0030 = 2·6374$

SUPPLEMENT III. MEAN DEVIATION AND MEAN DIFFERENCE.* LORENZ' AND PARETO'S CURVES.

(See pp. 114, 346.)

Let $y = f(x)$ be the equation of a continuous frequency curve, ranging from $x = h$ to $x = k$.

Write $Y = F(x) = \int_x^k f(x)dx$, *i.e.* the number of observations above x. Then N, the whole number of observations $= F(h)$.
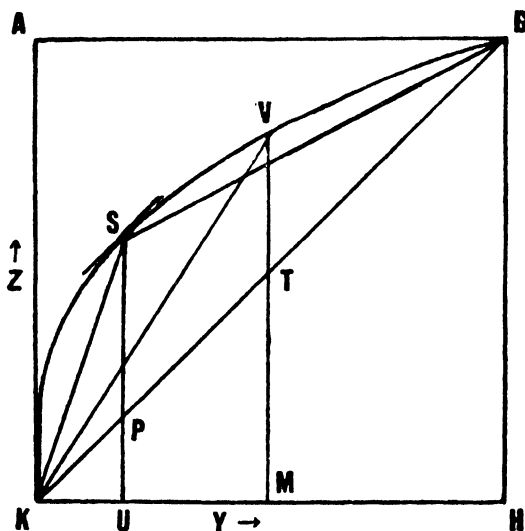
DIAGRAM C.

Write $Z = \Phi(x) = \int_x^k xf(x)dx$, *i.e.* the sum of the values of x from x to k. Then, if $\bar{x}$ is the average of the curve, the total of all the values of $x = N\bar{x} = \Phi(h)$.

We have $\dfrac{dY}{dx} = \dfrac{dF}{dx} = -f(x) = -y$

$$\frac{dZ}{dx} = \frac{d\Phi}{dx} = -xf(x) = -xy$$

$$\frac{dZ}{dY} = x \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (i)$$

---

* For a general discussion on the use of these quantities and most of the formulæ in this note, see *Bulletin de l'Institut International de Statistique*, Tome XXV, 3ᵐᵉ Livraison, 1931, pp. 189–320.

Write η_u for the *mean deviation* of the curve about $x = u$.

$$N\eta_u = \int_u^k (x - u)y\,dx + \int_h^u (u - x)y\,dx$$

$$= \Phi(u) - uF(u) + u(N - F(u)) - (N\bar{x} - \Phi(u))$$

$$= 2\{\Phi(u) - uF(u)\} + N(u - \bar{x}).$$

This is a minimum for the value of u which satisfies

$$2\,\Phi'(u) - 2uF'(u) - 2F(u) + N = 0,$$

i.e. when $-2uf(u) + 2uf(u) - 2F(u) + N = 0,$

i.e., when $F(u) = \dfrac{N}{2}$, and u is the median, say m.

Then $N\eta_m = 2\Phi(m) - N\bar{x} = 2\{\Phi(m) - \bar{x}F(m)\}$ . . . (ii)

If $u = \bar{x}$, $N\eta_{\bar{x}} = 2_{\bar{x}}\{\Phi(\bar{x}) - \bar{x}F(\bar{x})\}$ . . . . . . (iii)

We have $\eta_{\bar{x}} > \eta_m$, since the latter is the minimum value.

Mean Difference. The number of differences is $\frac{1}{2}N(N - 1)$, which may be taken as $\frac{1}{2}N^2$ for purposes of integration.

Write g for the mean difference of the observations.

$$\tfrac{1}{2}N^2g = \int_h^k f(x)\left\{\int_x^k f(u)\,(u - x)\,du\right\}dx,$$

for the difference $u - x$ taken positively is to be multiplied by $f(u)$ and $f(x)$, the number of observations at u and at x. We first keep $f(x)$ constant, and obtain the sum of the differences given by all ordinates whose abscissa is greater than x, and then integrate for x between extreme values.

$$\therefore \tfrac{1}{2}N^2g = \int_h^k f(x)\{\Phi(x) - xF(x)\}dx$$

$$= \int_{x=h}^{x=k} (Z\,dY - Y\,dZ) = 2\int_h^k Z\,dY - [ZY]_h^k$$

(integrating by parts)

$$= 2\int_0^N Z\,dY - N\bar{x}N,$$

since $Z = N\bar{x}$, $Y = N$, when $x = h$, and $Z = 0 = Y$ when $x = k$.

Draw a graph of Z as a function of Y, $KSVB$, $KH = AB = N$, $KA = HB = N\bar{x}$.

When $x = h$, $Z = HB$, $Y = KH$; as x increases to k, Z and Y diminish to zero at K.

Then $\tfrac{1}{2}gN^2 = 2\,\text{Area } KVBH - KABH$

$$\therefore \qquad \frac{g}{\bar{x}} = \frac{4KVB}{KABH},$$

where KVB is the curvilinear area bounded by KVB, KB. KM, KU are the values of Y at the median and average.

Verticals through M and U meet KB at T and P, and the curve at V and S.

Since $HB = KH \times \bar{x}$,

$$MT = KM \times \bar{x} = \bar{x}F(m)$$
$$UP = KU \times \bar{x} = \bar{x}F(\bar{x})$$

also $US = \Phi(\bar{x}), MV = \Phi(m)$

$\therefore N\eta_m = 2(MV - MT) = 2VT$, from equation (ii),

and $\qquad \dfrac{\eta_m}{\bar{x}} = \dfrac{4KVB}{KABH}$, where KVB is rectilinear.

Similarly from equation (iii)

$$\frac{\eta_{\bar{x}}}{\bar{x}} = \frac{4KSB}{KABH}, \text{ where } KSB \text{ is rectilinear.}$$

At S, $\dfrac{dZ}{dY} = \bar{x}$, from equation (i), and therefore the tangent at S is parallel to KB, and PS is the maximum ordinate of the figure $KVBPK$.

Without loss of generality we can choose our scales so that $N = 1$, $\bar{x} = 1$.

Then g, $\eta_{\bar{x}}$, η_m are identified with four times the area KSB (curvilinear), KSB (rectilinear) and KVB, which are (except in extreme cases of equality) in descending order of magnitude.

If, for example, x stands for income, any position, such as V, shows the proportion of aggregate income $(VM \div KA)$, that accrues to the proportion $(KM \div KH)$ of all holders of income over h.

If all incomes tended to be equal, the curve KVB would approximate to the line KB.

Write $\lambda = \dfrac{KSVB}{KABH} = \frac{1}{4} . \dfrac{g}{\bar{x}}$. Then λ is the Lorenz measurement of inequality of distribution, or (as it is sometimes termed) of *concentration* of income among the richer. With $\lambda = 0$ all incomes are equal; as λ approaches its maximum, $\frac{1}{4}$, a greater and greater proportion of income tends to be in the hands of the richest.

Application to Pareto's Curve.

Take k as infinite.

Y, the number of incomes above $x(\pounds) = \dfrac{A}{x^a}$, where $a > 1$, and A and α are constants.

$$N = A \div h^{\alpha}, \text{ and } Y = N\left(\frac{h}{x}\right)^{\alpha}.$$

$$Z = \int_x^{\infty} -x \cdot \frac{dY}{dx} \cdot dx = Nh^{\alpha}\int_x^{\infty}(+\alpha x^{-\alpha})dx = Nh^{\alpha} \cdot \alpha \cdot x^{1-\alpha} \div (\alpha - 1).$$

Write h for x, and we have $N\bar{x} = Nh\alpha \div (\alpha - 1)$.

$$\therefore \quad Z = N\bar{x}\left(\frac{h}{x}\right)^{\alpha-1}$$

Write $Y_1 = Y \div N$, the relative number of incomes above x, and $Z_1 = Z \div N\bar{x}$, the relative amount of income above x.

Then $Y_1 = \left(\frac{h}{x}\right)^{\alpha}$, $Z_1 = \left(\frac{h}{x}\right)^{\alpha-1}$,

and $Z_1 = Y_1^{1-\frac{1}{\alpha}} = Y_1^{\frac{1}{\delta}}$, if $\frac{1}{\alpha} + \frac{1}{\delta} = 1.*$

When α approaches 1, δ increases indefinitely and income is "concentrated" in few hands; at the same time inequality of incomes diminishes.

With this equation

$$\tfrac{1}{2}N^2 g = 2\int_0^1 N^2\bar{x}Z_1 \cdot dY_1 - N^2\bar{x} = N^2\bar{x}\left(\frac{2}{2-\frac{1}{\alpha}}-1\right)$$

$$g = \bar{x} \cdot \frac{2}{2\alpha - 1} = 2\bar{x} \cdot \frac{\delta - 1}{\delta + 1}.$$

$$\lambda = \frac{1}{2(2\alpha - 1)}$$

$$N\eta_{\bar{x}} = 2(Z_{\bar{x}} - \bar{x}Y_{\bar{x}}) = 2N\bar{x}\left\{\left(\frac{h}{\bar{x}}\right)^{\alpha-1} - \left(\frac{h}{\bar{x}}\right)^{\alpha}\right\}$$

$$\eta_{\bar{x}} = \bar{x} \cdot \frac{2(\alpha - 1)^{\alpha-1}}{\alpha^{\alpha}}$$

The median is given by $Y = \tfrac{1}{2}N$, $m = h \times 2^{\frac{1}{\alpha}}$.

$$\eta_m = \frac{2}{N}(Z_m - \bar{x}Y_m) = \bar{x}(2^{\frac{1}{\alpha}} - 1).$$

In the case where $\alpha = 1.5$,† we find

$\bar{x} = 3h$, $Y_1 = Z_1^3$, $g = 3h$, $\eta_{\bar{x}} = \frac{4h}{\sqrt{3}} = 2.31h$, $\eta_m = 1.76h$, $\delta = 3$, $\lambda = \tfrac{1}{4}$.

---

* The quantity δ in this connection is used by Professor Gini.
† The diagram illustrates this case.

Application to Normal Curve,

Write $y = f(x) = \dfrac{N}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}}$, k is ∞, h is $-\infty$, and $f(k) = 0 = f(h)$.

Then $\dfrac{dy}{dx} = -\dfrac{(x-\bar{x})}{\sigma^2}$. y, and since $\dfrac{dY}{dx} = -y$, we have

$(x - \bar{x}) \dfrac{dY}{dy} = \sigma^2$.

Now $\frac{1}{2} N^2 g = \displaystyle\int_h^k (Z dY - Y dZ) = N^2\bar{x} - 2\int_h^k Y dZ$

$$= N^2\bar{x} - 2\int_h^k xY dY$$

$$= N^2\bar{x} - 2\bar{x}\int_h^k Y dY - 2\int_h^k (x - \bar{x})Y dY$$

$$= 2\int_h^k (x - \bar{x})Y dY\,^*, \text{ since } 2\int_h^k Y dY = [Y^2]_h^k = N^2$$

* So far, true for any curve.

$$= 2\sigma^2\int_h^k Y dy = [2\sigma^2 y Y]_h^k - 2\sigma^2\int_h^k (-y^2)dx$$

$$= 0 + 2\sigma^2\int_h^k \frac{N^2}{2\pi\sigma^2} e^{-\frac{(x-\bar{x})^2}{\sigma^2}} dx$$

$$= \frac{\sigma N^2}{\sqrt{\pi}}$$

$$\therefore \qquad g = \frac{2\sigma}{\sqrt{\pi}}$$

We already know that the mean deviation $= \dfrac{\sigma\sqrt{2}}{\sqrt{.\pi}}$ (p. 269 (24)). This is consistent with equation (iii) p. 461, for $\bar{x} = m$, and $N\eta_m = N\eta_{\bar{z}}$

$$= 2(\Phi(\bar{x}) - \bar{x}\,F(\bar{x})) = 2\int_{\bar{z}}^{\infty} (x - \bar{x})\frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}} dx$$

$$= \frac{N}{\sqrt{\pi}} \sigma\sqrt{2} \left[e^{-\frac{(x-\bar{x})^2}{2\sigma^2}}\right]_{\bar{z}}^{\infty} = N . \frac{\sigma\sqrt{2}}{\sqrt{\pi}}$$

SUPPLEMENT IV.—TIME SERIES.

(See pp. 132, 137, 148, 337 *seq.*)

Suppose that we have a series of annual records which we have reason to think depend on a non-periodic trend, affected by sporadic variation that is independent of the date.

We may write for any observed value $Y = f(T) + v$, where $f(T)$ expresses the trend and v is the residual, T measuring the number of years from any zero time.

Take the particular case where $f(T)$ is expansible in a rapidly converging series and write

$$Y = a + bT + cT^2 + dT^3 + \ldots + v.$$

Transform this so that time (t) is measured from the centre of the period under consideration,

$$Y = a + bt + ct^2 + dt^3 + \ldots + v.$$

Write T_s for mean t^s.

Then when s is odd, $T_s = 0$. *I.e.* $T_1 = T_3 = T_5 = \ldots = 0$.

It can be shown, or verified, that, when n is the number of years considered,

$T_2 = (n^2 - 1)/12$, $T_4 = (n^2 - 1)(3n^2 - 7)/240$,

$$T_6 = T_2 . (3n^4 - 18n^2 + 31)/112,$$

$T_4 - T_2^2 = (n^2 - 1)(n^2 - 4)/180,$

$$T_2 T_6 - T_4^2 = (n^2 - 1)^2(n^2 - 4)(n^2 - 9)/33,600.$$

These formulæ are true whether n is odd or even.

Compute $a, b, c, d \ldots$ by the hypotheses that $0 = \text{mean}$ $v = \text{mean } vt = \text{mean } vt^2 = \ldots$, thus expressing arbitrarily the mutual independence of v and t.

Write $\bar{y}, m_1, m_2, m_3 \ldots$ for mean $Y, yt, yt^2, yt^3 \ldots$, where $y = Y - \bar{y}$.

Consider only powers up to t^3.
We have

$$Y = a + bt + ct^2 + dt^3 + v.$$

Add the n expressions of this kind and take the mean.

$\bar{y} = a + 0 + cT_2 + 0 + 0$, since mean $v = 0$

$\therefore y = Y - \bar{y} = bt + c(t^2 - T_2) + dt^3 + v.$

Multiply the last equation by t, and take the mean.

$m_1 = bT_2 + c(T_3 - T_2 . T_1) + dT_4 + 0$

$\quad = bT_2 + dT_4$, since mean $vt = 0$.

H H

Similarly multiply the same equation by t^2, and take the mean.

$$m_2 = c(T_4 - T_2{}^2), \text{ since mean } vt^2 = 0.$$

and again

$$m_3 = bT_4 + dT_6, \text{ since mean } vt^3 = 0.$$

Hence

$$Y = \bar{y} - \frac{m_2 T_2}{T_4 - T_2{}^2} + \frac{m_1 T_6 - m_3 T_4}{T_2 T_6 - T_4{}^2} \cdot t + \frac{m_2}{T_4 - T_2{}^2} t^2 + \frac{m_3 T_2 - m_1 T_4}{T_2 T_6 - T_4{}^2} t^3 + v.$$

The values of $\bar{y}$, $m_2 \ldots$ are to be computed from the observations. Those of $T_2, T_4 \ldots$ are given above in terms of n.

When d is zero, the equation becomes that of the parabola

$$Y = \bar{y} - \frac{15 m_2}{n^2 - 4} + \frac{12 m_1}{n^2 - 1} \cdot t + \frac{180 m_2}{(n^2 - 1)(n^2 - 4)} \cdot t^2 + v.$$

When c is also zero, m_2 is zero, and the equation is linear,

$$Y = \bar{y} + \frac{12 m_1}{n^2 - 1} \cdot t + v.*$$

In this last case $y = \dfrac{12 m_1}{n^2 - 1} \cdot t + v.$

Multiply by y and take the mean ı

$$\sigma_y{}^2 = \frac{12 m_1}{n^2 - 1} \cdot m_1 + \text{mean } vy.$$

Multiply by v and take the mean

$$\text{Mean } vy = 0 + \sigma_v{}^2$$

$$\therefore \sigma_v{}^2 = \sigma_y{}^2 - \frac{12 m_1{}^2}{n^2 - 1} = \sigma_y{}^2 - b^2 T_2 = \sigma_y{}^2 - \frac{m_1{}^2}{T_2}.$$

This expression indicates the reduction of the deviations when the observations are measured from the chosen linear trend. There is no improvement, if $0 = m_1 = \text{mean } yt$. At the other extreme $\sigma_v = 0$, if $y = bt$ throughout the period, for then $\sigma_y{}^2 = b^2 T_2$.

In the case of a parabola, we find similarly

$$\sigma_v{}^2 = \sigma_y{}^2 - b^2 T_2 - c^2(T_4 - T_2{}^2) = \sigma_y{}^2 - \frac{m_1{}^2}{T_2} - \frac{m_2{}^2}{T_4 - T_2{}^2}.$$

---

* It is easily seen that the coefficient of t is $r_{yt}\sigma_y/\sigma_t$ as in the usual regression equation.

For the cubic,

$$\sigma_r^2 = \sigma_y^2 - \frac{m_1^2}{T_2} - \frac{m_2^2}{T_4 - T_2^2} - \frac{(T_2 m_3 - T_4 m_1)^2}{T_2(T_2 T_6 - T_4^2)}$$

A measure of the significance of b and c (in the case of a parabola) may be determined thus :—

If there were no trend and the y's were distributed at random about zero, the value of $b = \Sigma yt/nT_2$ would be fortuitous.

Then b is a weighted sum of the y's, the weight of y_t being t/nT_2.

$$\therefore \quad \sigma_b^2 = \frac{\sigma_y^2 \Sigma t^2}{n^2 T_2^2} = \frac{\sigma_y^2}{nT_2} = \frac{12}{n(n^2-1)} \cdot \sigma_y^2 \text{ (p. 316).}$$

Similarly, $c = \Sigma yt^2/n(T_4 - T_2^2)$,

$$\sigma_c^2 = \sigma_y^2 \cdot \frac{nT_4}{n^2(T_4 - T_2^2)^2} = \sigma_y^2 \cdot \frac{135(3n^2 - 7)}{n(n^2-1)(n^2-4)^2}.$$

[See *Journal of the R.S.S.*, 1886, pp. 469–475, Edgeworth, and 1926, pp. 307, Bowley.]

Correlation of Time Series.

[*Journal of the R.S.S.*, 1926, pp. 300 *seq.*]

With the notation already used (p. 465), take two series

$$x = X - \bar{x} = b_1 t + c_1(t^2 - T_2) + u$$
$$y = Y - \bar{y} = b_2 t + c_2(t^2 - T_2) + v,$$

where b_1, b_2, c_1, c_2 are determined as before by the conditions $0 = \text{mean } u = \text{mean } v = \text{mean } ut = \text{mean } vt = \text{mean } ut^2 = \text{mean } vt^2$.

Multiply these equations and equate the means of the left- and right-hand products, remembering that $0 = T_1 = T_3$.

Mean $xy = b_1 b_2 T_2 + c_1 c_2(T_4 - T_2^2) + \text{mean } uv$.

$\therefore r_{xy} \cdot \sigma_x \cdot \sigma_y = b_1 b_2 T_2 + c_1 c_2(T_4 - T_2^2) + r_{uv} \cdot \sigma_u \sigma_v$.

Also by squaring each equation we have

$$\sigma_x^2 = b_1^2 T_2 + c_1^2(T_4 - T_2^2) + \sigma_u^2$$
$$\sigma_y^2 = b_2^2 T_2 + c_2^2(T_4 - T_2^2) + \sigma_v^2.$$

The equation for r_{xy} shows the contributions to a crude correlation coefficient, between two variables in time, made by the trend constants and by the residuals.

This can be better visualised in the case of linear trends, where $c_1 = c_2 = 0$.

Write $l_1 = b_1 \times \dfrac{n}{2} \div \sigma_u =$ the increase due to trend in half the range divided by the standard deviation of residuals from the trend. Similarly write $l_2 = b_2 \times \dfrac{n}{2} \div \sigma_v$.

Then $\quad r_{xy} = \{\tfrac{1}{3}l_1l_2\left(1 - \dfrac{1}{n^2}\right) + r_{uv}\} \cdot \sigma_u\sigma_v \, . \, /\sigma_x\sigma_y,$

since $T_2 = (n^2 - 1)/12$.

When n is great, this tends to

$$(\tfrac{1}{3}l_1l_2 + r_{uv})/\sqrt{\{(\tfrac{1}{3}l_1{}^2 + 1)(\tfrac{1}{3}l_2{}^2 + 1)\}}.$$

Thus when the trend-gradients are insignificant, r_{xy} is dominated by the correlation of the residuals, but when the gradients are considerable they outweigh the effect of the residuals.

Notice that if $b_2 = 0$, $r_{xy} = r_{uv}\sigma_u/\sigma_x$, and the trend of the x line does not affect the correlation.

SUPPLEMENT V.—THE LOGISTIC CURVE.

(Note to Chapter V, p. 343 *seq.*)

Let P measure population at any time t.

Then, if P increased in continual geometric progression, $\dfrac{1}{P} \cdot \dfrac{dP}{dt}$ would be constant, and P would become infinitely great as t increased without limit.

Empirical formulæ can be suggested to damp down the increase ; that best known is the Logistic Curve, whose equation is

$$\frac{1}{P} \cdot \frac{dP}{dt} = \frac{1}{a}\left(1 - \frac{P}{L}\right).$$

Here a and L are constants. Growth continues till P = L, which is the limit of population. a measures the time-scale.

The integral form is readily obtained as

$$P = L \div \left(1 + e^{\frac{b-t}{a}}\right)$$

where b is a constant.

Take the arbitrary time zero at b, writing $\tau = t - b$.

Then, $$P = L \div (1 + e^{-\frac{\tau}{a}}),$$

and $$P - \frac{L}{2} = \frac{L}{2} \cdot \frac{e^{\frac{\tau}{2a}} - e^{-\frac{\tau}{2a}}}{e^{\frac{\tau}{2a}} + e^{-\frac{\tau}{2a}}} = \frac{L}{2} \tan h \frac{\tau}{2a},$$

which is skew-symmetrical in τ about the value $P = \frac{L}{2}$.

The rate of growth is a maximum when

$$\frac{d^2P}{d\tau^2} = 0, \ \tau = 0, \ t = b.$$

A method [*] of evaluating L, a and b from the observations is to write $P = \frac{1}{Q}$.

Then $QL - 1 = e^{\frac{t-t}{a}}$.

Take three values, Q_{-1}, Q_0, Q_1 at equal time in intervals, viz. : $t = h$, o, $+ h$, where zero time is taken as b the middle observation; h may be a Census interval or a multiple thereof.

Then $Q_{-1}L - 1 = e^{\frac{b+h}{a}}$, $Q_0L - 1 = e^{\frac{b}{a}}$, $Q_1L - 1 = e^{\frac{b-h}{a}}$.

Then $\frac{h}{a} = \log_e (Q_0 - Q_{-1}) - \log_e (Q_1 - Q_0)$

$\frac{b}{a} = \log_e (Q_0 - Q_{-1}) + \log_e (Q_1 - Q_0) - \log_e (Q_1 Q_{-1} - Q_0^2)$

$L(Q_1 Q_{-1} - Q_0^2) = Q_1 + Q_{-1} - 2Q_0,$

as can easily be verified. Hence a, b and L can be found

A variant [†] of the equation is obtained by writing

$$\frac{1}{P} \cdot \frac{dP}{dt} = \frac{1}{a} \sqrt{\left(1 - \frac{P}{L}\right)}.$$

The integral of this may be written

$$P = \frac{4L}{e^{\frac{b-t}{a}} + e^{\frac{t-b}{a}} + 2} = L \operatorname{sech}^2 \frac{t - b}{2a}.$$

---

[*] See *R.S.S. Journal*, 1925, Yule, pp. 49, 50.
[†] This was suggested to me by Dr. Rhodes. It is a special form of an equation given by Dr. L. Hamburger. *Chemisch Weekblad*, Dec. 30, No. 5 (1933), Amsterdam. "Investigation on Complete Growth Functions," p. 121.

This curve is symmetrical about its maximum, $P = L$, when $t = b$.

As t increases positively or negatively, P tends asymptotically to zero. The population thus rises to a maximum and then falls towards zero.

An infinite number of other forms is possible, and the choice between them is arbitrary.

SUPPLEMENT VI.—TRANSFORMATIONS OF THE NORMAL CURVE.

(See p. 346.)

Suppose z to be a variable normally distributed with frequency

$$\eta = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

Let x be related to z by the equation $x = f(z)$, and consider the frequency distribution of x, say $y = F(x)$.

To every element ηdz at z, there will correspond an element $y dx$ at x. To $z = 0$, corresponds the median,* but not in general the average or mode, of $F(x)$.

The relationship between the curves of η and y will be as indicated in the diagram D annexed, where $f(z) = \frac{1}{5}(z + 5)^2$ is taken as an example. The rectangular area at N, where $z = 1$, becomes the broader and lower rectangle at N', where $x = 7 \cdot 2$.

Simple Case of Translation.

In Edgeworth's method of Translation $f(z)$ is taken in the form $x = a + bz + cz^2 + dz^3$, which is suitable when the variations of x depend on the variations of the cube of z. The full working out, with a suitable Table, is to be found in F. Y. Edgeworth's *Contributions to Mathematical Statistics*, issued as a separate pamphlet by the Royal Statistical Society in 1928. A simple case will illustrate the method.

Let $x = m + a(z + bz^2)$, where m is the median value of x.

This equation can be fitted to observation either by moments or by three percentiles.

* When formulæ in which $f''(z) \not> 0$ are excluded.

Method of Moments.

Measured from the median $M_t = \int_{-\infty}^{\infty} \{a(z + bz^2)\}^t \eta dz.$

$M_0 = 1$, $M_1 = ab$, $M_2 = a^2(1 + 3b^2)$, $M_3 = 3a^3b(3 + 5b^2)$, as may be found by direct integration.

I

II
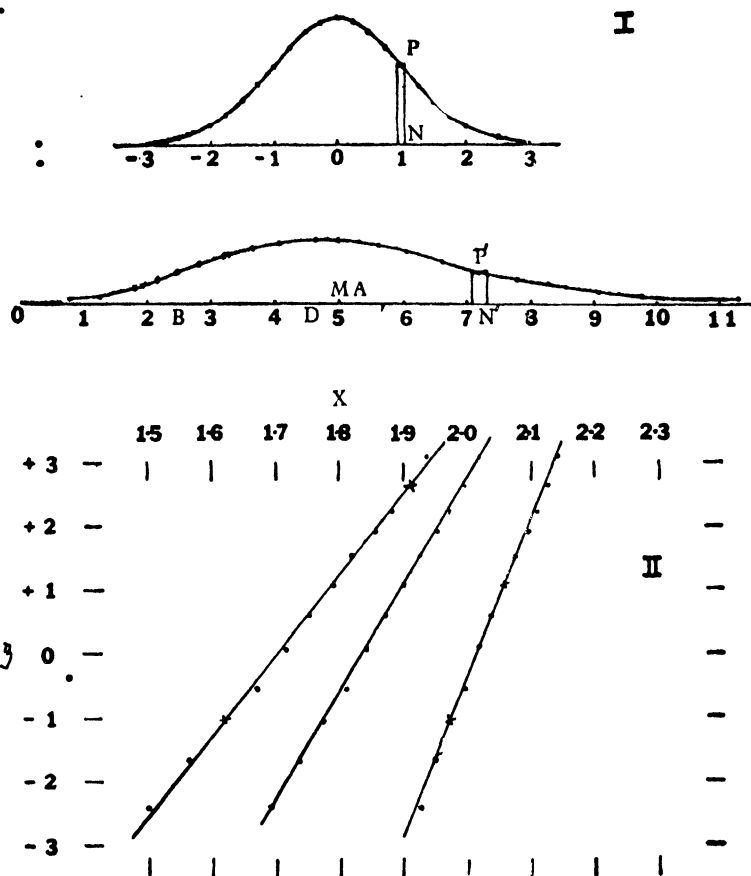
DIAGRAM D.

Therefore, measured from the average (see p. 251),

$$\mu_1 = 0, \mu_2 = a^2(1 + 2b^2), \mu_3 = 2a^3b(3 + 4b^2).$$

Write $2b^2 = w$, and as usual $\kappa = \mu_3 \div \mu_2^{\frac{3}{2}}$.

We readily find that

$$\kappa^2 = \frac{2w(3 + 2w^2)^2}{(1 + w)^3},$$

and therefore w is a root of

$$w^3(8 - \kappa^2) + 3w^2(8 - \kappa^2) + 3w(6 - \kappa^2) - \kappa^2 = 0$$

A first approximation gives $18w = \kappa^2$, which leads to

$$a = \sigma / \sqrt{1 + \frac{\kappa^2}{18}},$$

when $\sigma^2 = \mu_2$.

The full solution gives b, then the value of μ_2 gives a, and the median is $-ab$ from the average of the observations, $\bar{x}$,

so that $\qquad x = \bar{x} - ab + a(z + bz^2).$

Thus in the example on p. 305, $\kappa = \cdot 4093$, $w = \cdot 00943$, $b = \cdot 0687$, $\sigma = 9 \cdot 4155$, $a = \sigma \div \sqrt{1 \cdot 00943} = 9 \cdot 372$, $m = 51 \cdot 453 - ab = 50 \cdot 810$, and

$$x = 50 \cdot 810 + 9 \cdot 372(z + \cdot 0687z^2) \quad . \quad . \quad . \quad \text{(i)}$$

Method of percentiles.

Let z_1, z_2, z_3 correspond to x_1, x_2, x_3.

The known proportion (p_1) of these observations less than x_1 is the same as the proportion below z_1;

Hence $-\frac{1}{2} + p_1 = \int_0^{z_1} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$, and z can be written

down from the normal table.

We then have $x_1 = m + a(z_1 + bz_1^2)$ and two similar equations to determine m, a and b. b is found from

$$\frac{x_2 - x_1}{x_3 - x_2} = \frac{(z_2 - z_1)(1 + b(z_2 + z_1))}{(z_3 - z_2)(1 + b(z_3 + z_2))},$$

and thence a and m.

Select three values from the Table on p. 305.

$x_1 = 31 \cdot 5$, $p_1{}^* = \cdot 008$, $-\frac{1}{2} + p_1 = -\cdot 492$, $z_1 = -2 \cdot 410$
$x_2 = 51 \cdot 5$, $p_2 = \cdot 525$, $-\frac{1}{2} + p_1 = \cdot 025$, $z_2 = \cdot 063$
$x_3 = 86 \cdot 5$, $p_3 = \cdot 999$, $-\frac{1}{2} + p_1 = \cdot 499$, $z_3 = 3 \cdot 10$

The solution of the equations gives

$$x = 50 \cdot 85 + 9 \cdot 53(z + \cdot 0653z^2) \quad . \quad . \quad . \quad \text{(ii)}$$

* Obtained by adding the "observations" column to the given value of x.

The observations are compared with the curve by finding the values of z that correspond to the given values of x, and so the number that should fall in each grade. Or we can write down the values of z that correspond to successive groups of occupations and find how nearly the resulting x's agree with the limits of the grades in the data.

The method is only applicable when κ is not great, say not greater than unity (see *R.S.S. Journal*, 1898, pp. 695–7).

When κ is small, so that κ^2 is negligible, the distribution tends to that given on p. 302.

Within these limits the method is useful.

The Law of Proportional Effect.

Take the transforming equation as $x = x_0 + e^{\frac{z+b}{a}}$,

so that $z = a \log_e (x - x_0) - b$.

Then $dz = a \cdot \dfrac{d(x - x_0)}{x - x_0}$, so that a small absolute variation in z is proportional to a small relative variation in $(x - x_0)$.

The frequency of x is given by

$$y\,dx = \eta\,dz$$
$$y = \frac{a}{x - x_0} \cdot \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}\{a \log_e (x - x_0) - b\}^2} = F(x)$$

The constants x_0, a, and b are to be determined from the observations.

Diagram D. I will serve to indicate the relative position of the points, though it is drawn from a " translation " equation.

O is the origin from which the x observations are measured. M, D and A are respectively the median, mode and average of the curve $y = F(x)$. OB $= x_0$.

The median of the new curve corresponds to $z = 0$.

Hence OM $= x_0 + e^{\frac{b}{a}}$, and $\log_e$ BM $= \dfrac{b}{a}$,

The mode is obtained from $\dfrac{d}{dx} \log y = 0$, which leads to

$$\log_e \text{BD} = \frac{b}{a} - \frac{1}{a^2}.$$

The area remains unity and the average is given by

$$\bar{x} = \text{OA} = \int_{x_0}^{\infty} xy\,dx = \int_{-\infty}^{\infty} \left(x_0 + e^{\frac{b+z}{a}}\right)\eta\,dz$$

$$= x_0 + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(z-\frac{1}{a}\right)^2} \cdot e^{\frac{b}{a}+\frac{1}{2a^2}}\,dz = x_0 + e^{\frac{b}{a}+\frac{1}{2a^2}}$$

$$\therefore \log \text{BA} = \log(\text{OA} - x_0) = \frac{b}{a} + \frac{1}{2a^2}.$$

Hence $\log \text{BD} + 2\log \text{BA} = 3\log \text{BM}$, and $\log \dfrac{\text{BM}}{\text{BA}} = \tfrac{1}{3}\log \dfrac{\text{BD}}{\text{BA}}$, which is analogous to equation (145), p. 444.[*]

To find the constants we can again proceed either by the use of moments or of percentiles, or graphically.

Method of Moments. The average has already been found.

$$\mu_2 = \int_{-\infty}^{\infty} \left(e^{\frac{b+z}{a}} - e^{\frac{b}{a}+\frac{1}{2a^2}}\right)^2 \cdot \frac{e^{-\frac{1}{2}z^2}}{\sqrt{2\pi}}\,dz.$$

The integration is fairly simple and leads to

$$e^{\frac{2b}{a}+\frac{1}{a^2}}\left(e^{\frac{1}{a^2}} - 1\right).$$

Similarly

$$\mu_3 = e^{\frac{3b}{a}} \cdot \left(e^{\frac{9}{2a^2}} - 3e^{\frac{5}{2a^2}} + 2e^{\frac{3}{2a^2}}\right) = e^{\frac{3b}{a}+\frac{3}{2a^2}}\left(e^{\frac{1}{a^2}} - 1\right)^2\left(e^{\frac{1}{a^2}} + 2\right)$$

$$\therefore \qquad \kappa^2 = \mu_3{}^2/\mu_2{}^3 = \left(e^{\frac{1}{a^2}} + 2\right)^2\left(e^{\frac{1}{a^2}} - 1\right)[†]$$

These equations are sufficient to determine a, b and x_0.
We apply them to the example on p. 305.

Write $v = e^{\frac{b}{a}}$, $w = e^{\frac{1}{a^2}}$.
Then

$$(w + 2)(w - 1)^{\frac{1}{2}} = \kappa = \cdot 4093. \qquad \therefore w = 1 \cdot 0184,\ a = 7 \cdot 406$$

$$vw^{\frac{1}{2}}(w - 1)^{\frac{1}{2}} = \sigma = 9 \cdot 4155.$$

---

[*] In the translation method we have the result $\text{MA} = \frac{1}{3}\text{DA}$, when x^2 and the coefficient of x^3 are neglected.

[†] This is equivalent to an equation given by Wicksell, *Genetic Theory of Frequency*, 1917, p. 14, equation (22).

$v = 68 \cdot 78, b = a \log_e v = 31 \cdot 33.$

$x_0 + vw^t = \bar{x} + 51 \cdot 453. \quad x_0 = -17 \cdot 89.$

$z = 7 \cdot 406 \log_e (x + 17 \cdot 9) - 31 \cdot 33$

$\quad = 17 \cdot 05 \log_{10} (x + 17 \cdot 9) - 31 \cdot 33 \quad . \quad . \quad .$ (iii)

Method of Percentiles. For ease of solution we require to know the median and two percentiles at equal proportions from it. In general we only know one of these except by approximation.

Let p_1 be the proportion of the observations between the median and either chosen percentile, and x_1, m, x_2 be the scale readings for these percentiles and the median.

Determine $z = h$ from the Table on p. 271 so that

$$\int_0^h \frac{1}{\sqrt{2\pi}} e^{\frac{z^2}{2}} dz = p_1$$

Then $x_1 = x_0 + e^{\frac{b-h}{a}}$, $m = x_0 + e^{\frac{b}{a}}$, $x_2 = x_0 + e^{\frac{b+h}{a}}$.

In the same example (p. 305), the median by interpolation is $50 \cdot 945 = m$.

Take $x_1 = 41 \cdot 5$, whence $p_1 = \cdot 357$, and $h = 1 \cdot 067$.

For x_2, we must include $\cdot 857$ of the observations. We obtain by interpolation $x_2 = 61 \cdot 34$.

Write $\quad v = e^{\frac{b}{a}}, u = e^{\frac{h}{a}}$

$41 \cdot 5 \quad = x_1 = x_0 + v/u, \quad 50 \cdot 945 = m = x_0 + v,$

$61 \cdot 34 = x_2 + vu.$

$u = \dfrac{x_2 - m}{m - x_1} = 1 \cdot 10$, and $a = 11 \cdot 165.$

$\dfrac{1}{v} = \dfrac{1}{m - x_1} - \dfrac{1}{x_2 - m}$, and $v = 103 \cdot 4.$

$b = a \log_e v = 51 \cdot 77.$

$x_0 = 50 \cdot 945 - v = -52 \cdot 45.$

$z = 11 \cdot 165 \log_e (x - x_0) - 51 \cdot 77$

$\quad = 25 \cdot 7 \log_{10} (x + 52 \cdot 5) - 51 \cdot 8 \quad . \quad . \quad .$ (iv)

It is noticeable that this equation differs in its constants from that by the method of moments. But, as seen in the diagram on p. 471, the middle and right-hand lines both give a plausible fit to the observations. It is remarkable what little change in the result is made by great changes in the constants.

M. Gibrat's Graphic Method.*

From the observations we can find, by the method just used, the values of z that correspond to the given values of x.†

We have therefore a number of equations of the form $z = a \log_e (x - x_0) - b$, where z and x are known, and x_0, a, and b are unknown.

Take an arbitrary value of x_0, write $X = \log_e (x - x_0)$, so that $z = aX - b$. Plot z, X on squared paper from the observa-

DISTRIBUTION OF 1000 SUMS OF THE NUMBER OF LETTERS IN 10 WORDS.

1 x.	2 F(x).	3 z.	4 log x.	5 log x + 17.9	6 log x + 52.5	7 Data.	8 Proportional Effect iii.	9 iv.	10 v.	11 Trans-lation i.	12 Normal Second Approx.
						8	5	7	10	7	8
31·5	·492	− 2·41	1·50	1·69	1·92						
						38	34	35	33	41	37
36·5	·454	− 1·68	1·56	1·74	1·95						
						97	103	98	93	95	99
41·5	·357	− 1·07	1·62	1·77	1·97						
						155	184	172	172	170	173
46·5	·202	− 0·53	1·67	1·81	1·996						
						227	222	212	204	216	213
51·5	·025	+ 0·06	1·71	1·84	2·017						
						202	188	198	193	191	191
56·5	·227	+ 0·61	1·75	1·87	2·037						
						134	126	133	150	136	135
61·5	·361	+ 1·09	1·79	1·90	2·057						
						76	75	77	84	80	78
66·5	·437	+ 1·53	1·82	1·93	2·075						
						37	36	41	38	38	40
71·5	·474	+ 1·94	1·85	1·95	2·093						
						13	16	17	16	17	18
76·5	·487	+ 2·23	1·88	1·975	2·106						
						0	7	7	5	6	7
81·5	·496	+ 2·65	1·91	1·997	2·127						
						3	3	2	1	2	2
86·5	·499	+ 3·09	1·94	2·019	2·143						
						1	1	1	1	1	0
					χ^2		13	5½	14	6	7½

Col. 1 is the original scale. Col. 2 is computed from the Data, col. 7, and shows the proportion of observations from the centre to the value of x.　Col. 3 is from the Table of the Normal Curve.　Cols. 8 to 11 are the results of computing z from equations (iii), (iv), (v) and (i) respectively, reading the corresponding $F(z)$ from the normal Table and writing down the differences.　Col. 12 is from p. 305.

tions. If the points indicate concavity (or convexity) to the axis of Z, decrease (or increase) x_0 and re-plot the points. After one or two experiments a straight line will be found approximately, if in fact the observations can be represented

* R. Gibrat, *Les Inégalités Économiques*, Recueil Sirey, Paris, 1931. M. Gibrat introduces the term " La loi de l'effet proportionel," and gives many illustrations of its use.　The equation itself has been known for a long time. (Cf. Wicksell, *loc. cit.*).

† Thus in the Table annexed eight observations (Col. 7) are below 31·5 (Col. 1); therefore $x = 31·5$ marks ($\frac{1}{2} + ·492$) of the 1000 observations, or ·492 from the median (Col. 2).　(8 + 38) observations are below 36·5, which marks ·454 from the median.　Col. 3 is then obtained from the Table on p. 271, with due regard to sign.

by this form. From the graph, or by computation from two selected points, a and b can then be found:

Diagram D. II shows three lines on this basis. That on the left is

$$z = 11 \cdot 62 \log_{10} x - 19 \cdot 83 \quad . \quad . \quad . \quad . \quad (v),$$

the next is from equation (iii), that to the right is from equation (iv). The observations are shown by the dots. (v) indicates slight convexity to the axis of Z, (iv) slight convexity, (iii) is neutral.

The results, together with that by the moment method of translation are given in the Table, where also the last column of the Table on p. 305 (the result of the second approximation to the normal curve) is repeated.

Roughly computed values of χ^2 (p. 431) indicate that any of the equations give a plausible fit; (iii) and (iv) are the least satisfactory; (iv) is the best, closely followed by (i), but $P < \cdot 5$ also for the last column.

SUPPLEMENT VII.—CORRELATION OF RANKS.

(Note to pp. 368–9.)

Let n persons be arranged in order as regards each of two attributes, so that the tth person is of rank x_t for the one and y_t for the other. Then the values of x_t are $1, 2 \ldots n$, as are the values of y_t, but in general x_t is not the same as y_t.

Write $\bar{x}$ and $\bar{y}$ for the averages and σ_x, σ_y for the standard deviations.

$$\bar{x} = \frac{1}{n} \sum_{t}^{n} t = \frac{n+1}{2} = \bar{y}$$

$$\sigma_x^2 = \frac{1}{n} \sum_{t}^{n} t^2 - \left(\frac{n+1}{2}\right)^2 = \frac{n^2-1}{12} = \sigma_y^2$$

Write D^2 for the mean square difference between x_t and y_t.

$$D^2 = \frac{1}{n} \sum_{i}^{n} (x_t - y_t)^2 = \frac{1}{n} \Sigma\{(x_t - \bar{x}) - (y_t - \bar{y})\}^2$$

$$= \sigma_x^2 + \sigma_y^2 - 2R\sigma_x\sigma_y = \frac{1}{6}(n^2 - 1)(1 - R),$$

where R is computed as a correlation coefficient.

$$\therefore R = 1 - \frac{6D^2}{n^2 - 1} \text{ (Spearman's coefficient)}$$

is a measure of the difference between the rankings of the same person with respect to the two attributes.

$R = 1$ corresponds to identical orders, where $x_i = y_i$ for each person.

If the orders are reversed so that $x_i = n + 1 - y_i$,

$$\Sigma(x_i - y_i)^2 = \Sigma(2x_i - n - 1)^2$$
$$= 4\Sigma x_i^2 - 4(n + 1)\Sigma x_i + n(n + 1)^2$$
$$= \frac{n(n^2 - 1)}{3},$$

and then $R = -1$.

R does not measure the correlation between the attributes, which is not obtainable from the data, and the true correlation may have any number of values for unchanged R. For example, R depends only on the orders in which candidates in an examination are placed, and not on the actual marks.

Professor Karl Pearson has shown (*Drapers' Company Research Memoirs*, No. IV) that, when there are an indefinitely large number of persons with attributes distributed normally with coefficient of correlation r, we have

$$r = 2 \sin (\tfrac{1}{6}\pi R).$$

$r > R$, except when $r = R = 0$ or 1, but the maximum difference between r and R is only ·018, when $r = ·6$, approx.

SUPPLEMENT VIII.—NOTE ON DETERMINANTS. RECTILINEAR REGRESSION.

(Note to Chapter VIII.)

Since only one property of Determinants is necessary for the treatment of linear regression equations, it is worth while to obtain it from first principles.

Write $D = | a_1 \ b_2 \ c_3 \ d_4 | = \begin{vmatrix} a_1 & b_1 & c_1 & d_1 \\ a_2 & b_2 & c_2 & d_2 \\ a_3 & b_3 & c_3 & d_3 \\ a_4 & b_4 & c_4 & d_4 \end{vmatrix}$

$$= a_1\ b_2\ c_3\ d_4 - b_1\ a_2\ c_3\ d_4 + c_1\ a_2\ b_3\ d_4 - d_1\ a_2\ b_3\ c_4$$
$$- a_1\ b_2\ c_4\ d_3 + b_1\ a_2\ c_4\ d_3 - c_1\ a_2\ b_4\ d_3 + d_1\ a_2\ b_4\ c_3$$
$$+ a_1\ b_4\ c_2\ d_3 - b_1\ a_4\ c_2\ d_3 + c_1\ a_4\ b_2\ d_3 - d_1\ a_4\ b_2\ c_3$$
$$- a_1\ b_4\ c_3\ d_2 + b_1\ a_4\ c_3\ d_2 - c_1\ a_4\ b_3\ d_2 + d_1\ a_4\ b_3\ c_2$$
$$+ a_1\ b_3\ c_4\ d_2 - b_1\ a_3\ c_4\ d_2 + c_1\ a_3\ b_4\ d_2 - d_1\ a_3\ b_4\ c_2$$
$$- a_1\ b_3\ c_2\ d_4 + b_1\ a_3\ c_2\ d_4 - c_1\ a_3\ b_2\ d_4 + d_1\ a_3\ b_2\ c_4$$

Here every possible permutation of 1, 2, 3, 4 is used in the order of suffixes applied to a, b, c, d. One interchange of adjacent suffixes (or letters) is taken as changing the sign from $+$ to $-$, or from $-$ to $+$. The first term being taken as positive, the sign of every other term is determined.

Collect the coefficients (or " co-factors ") of a_1, b_1, c_1, d_1, and write them as D_{11}, D_{12}, D_{13}, D_{14} respectively.

Then $D = a_1 D_{11} + b_1 D_{12} + c_1 D_{13} + d_1 D_{14}$

Now, if the quantities in two rows are coincident, so that, for example, $a_1 = a_2$, $b_1 = b_2$, $c_1 = c_2$, $d_1 = d_2$, the rule of signs at once gives $D = 0$.

Therefore we have $0 = a_2 D_{11} + b_2 D_{12} + c_2 D_{13} + d_2 D_{14}$.

The co-factors are evidently determinants with one row and one column fewer than in the original determinant.

These definitions are easily generalised. Write $D = |a_{11}\ a_{22}$ $\ldots a_{nn}|$, where the first suffix determines the row, the second the column.

Then $a_{11} D_{11} + a_{12} D_{12} + \ldots + a_{1n} D_{1n} = D$ . . (a)

$$\left. \begin{array}{l} a_{21} D_{11} + a_{22} D_{12} + \ldots + a_{2n} D_{1n} = 0 \\ a_{31} D_{11} + a_{32} D_{12} + \ldots + a_{3n} D_{1n} = 0 \\ \quad\vdots \\ a_{n1} D_{11} + a_{n2} D_{12} + \ldots + a_{nn} D_{1n} = 0 \end{array} \right\} \quad (\beta)$$

For example, in the determinant $D = \begin{vmatrix} a & h & g \\ h & b & f \\ g & f & c \end{vmatrix}$,

$$a (bc - f^2) + h(fg - ch) + g(hf - bg) = D$$
$$h(bc - f^2) + b(fg - ch) + f(hf - bg) = 0$$
$$g(bc - f^2) + f(fg - ch) + c(hf - bg) = 0$$

Rectilinear Regression. 2 Variables.

Write
$$Y = aX + b + v \quad . \quad . \quad . \quad . \quad . \quad . \quad (i)$$
and $Y = y + \bar{y}$, $X = x + \bar{x}$, where $\bar{y}$ and $\bar{x}$ are the averages of the variables y, x.

Assume that mean $v = 0 =$ mean vx, thus expressing independence of v and x. v is then the residual, when Y is computed from X.

The mean of (i) is
$$\bar{y} = a\bar{x} + b + 0, \text{ since mean } v = 0 \quad . \quad . \quad . \text{ (ii)}$$
Subtract (ii) from (i).
$$y = ax + v \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \text{ (iii)}$$
Multiply (iii) by x and take the mean.
$$\text{mean } xy = a\sigma_x{}^2 + 0, \text{ since mean } vx = 0.$$
$$\therefore r\sigma_y = a\sigma_x \quad . \quad . \quad . \quad . \quad . \quad . \quad . \text{ (iv)}$$

Here σ_x, σ_y are the standard deviations of x, y and r is their correlation coefficient.

Multiply (iii) by v and take the mean.
$$\text{mean } vy = 0 + \sigma_v{}^2 \quad . \quad . \quad . \quad . \quad . \quad . \text{ (v)}$$
Multiply (iii) by y and take the mean.
$$\sigma_y{}^2 = ar\sigma_x\sigma_y + \text{mean } vy.$$
$$\therefore \sigma_y{}^2 = r^2\sigma_y{}^2 + \sigma_v{}^2, \text{ from (iv) and (v).}$$
$$\text{and } \sigma_v = \sigma_y \sqrt{(1 - r^2)}.$$

Equation (i) becomes $Y = \bar{y} + r \cdot \dfrac{\sigma_y}{\sigma_x} (X - \bar{x}) + v.$

The reduction of the standard deviation from σ_y to σ_v is one indication of how far our knowledge of Y is improved by estimating it with the help of X as $\bar{y} + r\dfrac{\sigma_y}{\sigma_x} (X - \bar{x})$.

For example, in a collection of 145 budgets of family expenditure, the average whole weekly expenditure per equivalent adult was found to be $16·9 = \bar{x}$, with $\sigma_x = 7·20$ (shillings). The expenditure on meat was $2·36 = \bar{y}$, when $\sigma_y = 1·14$.

r was found to be ·56.

The regression equation—meat on total expenditure—is therefore

$$Y = 2·36 + ·56 \times \frac{1·14}{7·20} (X - 16·9) + v.$$
$$= ·09 X + ·87 + v.$$
$$\sigma_v{}^2 = 1·14 \sqrt{(1 - ·56^2)} = ·95.$$

The reduction in standard deviation is not great, but the distribution of v is less unsymmetrical and more nearly normal than that of y.

Distributions of Y and of v. Compared with Normal Distribution.

Between		Normal.	Y +	Y −	v +	v −
o and	$\pm \frac{1}{3}\sigma$	18·9	20	17	20	29
$\pm \frac{1}{3}\sigma$	$\pm \frac{2}{3}\sigma$	16·9	8	27	17	20
$\pm \frac{2}{3}\sigma$	$\pm \sigma$	13·7	9	21	12	9
$\pm \sigma$	$\pm \frac{4}{3}\sigma$	9·8	10	10	8	8
$\pm \frac{4}{3}\sigma$	$\pm \frac{5}{3}\sigma$	6·3	4	6	4	3
$\pm \frac{5}{3}\sigma$	$\pm 2\sigma$	3·6	6	0	3	1
$\pm 2\sigma$	$\pm 3\sigma$	3·1	5	1	4	6
$\pm 3\sigma$	∞	0·2	1	0	1	0
		72·5	63	82	69	76

[Allen and Bowley : *Family Expenditure*, p. 82.]

Rectilinear Regression. n *Variables.*

Let a variable x_1 be related to variables $x_2, x_3 \ldots x_n$, all measured from their averages, by the equation

$$- x_1 = a_2 x_2 + a_3 x_3 + \ldots + a_n x_n + v, \quad \ldots \ldots \text{(i)}$$

where v is a residual such that $0 = \text{mean } vx_2$

$$= \text{mean } vx_3 \ldots = \text{mean } vx_n.*$$

Multiply (i) by x_1 and take the mean.

$$- \sigma_1{}^2 = a_2 r_{12}\sigma_1\sigma_2 + a_3 r_{13}\sigma_1\sigma_3 + \ldots + a_n r_{1n}\sigma_1\sigma_n + \text{mean } vx_1 \text{ (ii)}$$

where σ_1, σ_2 are the standard deviations of $x_1, x_2 \ldots$, and r_{12} is the coefficient of correlation between x_1, x_2 and so on. Of course, $r_{12} = r_{21}$, etc.

Multiply (i) by v and take the mean.

$$- \text{mean } vx_1 = 0 + 0 + \ldots + 0 + \sigma_v{}^2 \quad \ldots \ldots \text{(iii)}$$

Multiply (i) by x_2 and take the mean.

$$- r_{12}\sigma_1\sigma_2 = a_2\sigma_2{}^2 + a_3 r_{23}\sigma_2\sigma_3 + \ldots + a_n r_{2n}\sigma_2\sigma_n + 0$$

$$\therefore r_{21} \cdot \sigma_1 + a_2\sigma_2 + a_3\sigma_3 r_{23} + \ldots + a_n\sigma_n r_{2n} = 0$$

Similarly

$$r_{31} \cdot \sigma_1 + a_2\sigma_2 r_{32} + a_3\sigma_3 + \ldots + a_n\sigma_n r_{3n} = 0$$

$$\vdots$$

$$r_{n1}\sigma_1 + a_2\sigma_2 r_{n2} + a_3\sigma_3 r_{n3} + \ldots + a_n\sigma_n = 0$$

$$\left. \right\} \quad \ldots \text{(iv)}$$

and from (ii)

* These conditions are equivalent to those obtained by the Method of Least Squares, where Σv^2 is minimised (see p. 452–4).

I I

$$\sigma_1 + a_2\sigma_2 r_{12} + a_3\sigma_3 r_{13} + \ldots + a_n\sigma_n r_{1n} = -\frac{1}{\sigma_1} \text{ mean } vx_1$$

$$= \frac{1}{\sigma_1} \cdot \sigma_v{}^2, \text{ from (iii)} \quad \ldots \quad \ldots \quad \ldots \quad \text{(v)}$$

Write D † $= \begin{vmatrix} 1 & r_{12} & r_{13} \ldots r_{1n} \\ r_{21} & 1 & r_{23} \ldots r_{2n} \\ \vdots \\ r_{n1} & r_{n2} & r_{n3} \ldots 1 \end{vmatrix}$

† Written R on p. 407.

Then the values of a_2, a_3 . . . are given by

$$\frac{\sigma_1}{D_{11}} = \frac{a_2\sigma_2}{D_{12}} = \frac{a_3\sigma_3}{D_{13}} = \ldots = \frac{a_n\,\sigma_n}{D_{1n}},$$

for if these are substituted in equation (iv), we have

$$r_{21} \cdot D_{11} + 1 \cdot D_{12} + r_{23}D_{13} + \ldots r_{2n}D_{1n} = 0$$
$$r_{31} \cdot D_{11} + r_{32} \cdot D_{12} + D_{13} + \ldots r_{3n}D_{1n} = 0$$
$$\vdots$$
$$r_{n1} \cdot D_{11} + r_{n2} \cdot D_{12} + r_{n3} \cdot D_{13} + \ldots 1 \cdot D_{1n} = 0$$

which are identically true from equations (β), p. 479.

Equation (i) may now be written

$$x_1 = -\frac{\sigma_1}{D_{11}}\left(\frac{x_2}{\sigma_2}D_{12} + \frac{x_3}{\sigma_3}D_{13} + \ldots + \frac{x_n}{\sigma_n}D_{1n}\right) + v$$

(p. 408, last line, where $R_{12} = D_{12}$, etc.).

Also from (v),

$$\frac{1}{\sigma_1}\sigma_v{}^2 = \frac{\sigma_1}{D_{11}}\left(1 \cdot D_{11} + r_{12}D_{12} + r_{13}D_{13} + \ldots + r_{1n}D_{1n}\right)$$

$$= \frac{\sigma_1}{D_{11}} \cdot D, \text{ from equation } (a) \text{ p. 479.}$$

$$\therefore \sigma_v{}^2 = \sigma_1{}^2 \cdot \frac{D}{D_{11}}.$$

When $n = 3$, $D = 1 + 2r_{23}r_{31}r_{12} - r_{23}{}^2 - r_{31}{}^2 - r_{12}{}^2$

$$= 1 \cdot (1 - r_{23}{}^2) + r_{12}(r_{23}r_{31} - r_{12})$$
$$+ r_{13}(r_{12}r_{23} - r_{13})$$

$$D_{11} = 1 - r_{23}{}^2, \quad D_{12} = r_{23}r_{31} - r_{12}, \quad D_{13} = r_{12}r_{23} - r_{13}$$

$$x_1 = \frac{r_{12} - r_{23}r_{31}}{1 - r_{23}{}^2} \cdot \frac{\sigma_1}{\sigma_2}x_2 + \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}{}^2} \cdot \frac{\sigma_1}{\sigma_3} \cdot x_3 + v$$

(p. 400 (115)).

$$\sigma_v{}^2 = \sigma_1{}^2(1 + 2r_{23}r_{31}r_{12} - r_{23}{}^2 - r_{31}{}^2 - r_{12}{}^2)/(1 - r_{23}{}^2).$$

When $n = 2$, $D = 1 - r_{12}{}^2 = 1 + r_{12}(-r_{12})$, $D_{11} = 1$, $D_{12} = -r_{12}$

$$x_1 = \frac{r_{12}\sigma_1}{\sigma_2}x_2 + v, \quad \sigma_v{}^2 = \sigma_1{}^2(1 - r_{12}{}^2) \div 1$$

as before.

SUPPLEMENT IX.—FREQUENCY OF THE SECOND MOMENT IN SMALL SAMPLES.

(See pp. 416–421.)

Great importance is attached in modern work to equation (i) in the following, and it is well to show its relationship to the formulæ on pages 417, 420, 421 and 451.

Let $x_1, x_2 \ldots x_n$ be n quantities selected at random from a normal group whose average and second moment are zero and m_2.

Write $\overset{n}{\underset{t=1}{S}} x_t = n\bar{x}$, and $Sx_t^2 - n\bar{x}^2 = n\mu_2$.

Required the chance of obtaining $\bar{x}$, μ_2.

The chance that $\ldots x_t$ to $x_t + dx_t \ldots$ should be drawn is

$$(2\pi m_2)^{-\frac{n}{2}} \cdot e^{-\frac{1}{2m_2}Sx_t^2} \ldots dx_t \ldots$$

$$= (2\pi m_2)^{-\frac{n}{2}} e^{-\frac{n\bar{x}^2}{2m_2}} \cdot e^{-\frac{n}{2m_2}\mu_2} \ldots dx_t \ldots$$

$$= C \cdot F(\bar{x}) \cdot e^{-\frac{n}{2m_2}\mu_2} dx_1 \ldots dx_t \ldots dx_n,$$

Here $F(\bar{x})$ is the normal error function with standard deviation $\sqrt{m_2/n}$ (p. 290 (41)). μ_2, and therefore $\frac{n}{m_2}\mu_2$ is a homogeneous quadratic function of n quantities, or, if $\bar{x}$ is given, of $n - 1$ quantities, and is analogous to X^2 in p. 493 below. Using the same transformation as there, we have

Chance of obtaining μ_2 with a *given* $\bar{x}$ is

$$K \cdot F(\bar{x}) e^{-\frac{n}{2m_2}\mu_2} (\sqrt{\mu_2})^{n-2} d\sqrt{\mu_2} = K_1 \cdot F(\bar{x}) e^{-\frac{n}{2m_2}\mu_2} \mu_2^{\frac{n-3}{2}} d\mu_2,$$ where K and K_1 are constants.

Hence the chance of obtaining μ_2 for *any* $\bar{x}$ is found by integrating $F(\bar{x})$ from $-\infty$ to $+\infty$, and is

$$P = K_2 \cdot e^{-\frac{n}{2m_2}\mu_2} \cdot \mu_2^{\frac{n-3}{2}} d\mu_2,$$ where K_2 is constant . . . (i)

[Here $\frac{1}{P} \cdot \frac{dP}{d\mu_2} = \frac{-n\mu_2 + (n-3)m_2}{2m_2\mu_2}$, which is Pearson's Type III. Cf. page 344 (84), with $b_0 = b_2 = 0, y = P, x = \mu_2, b_1 = -2m_2/n, a = -(n-3)m_2/n.$]

Write $\frac{n}{2m_2}\mu_2 = u$, and $n = 2p + 1$.

$$P = K_3 e^{-u} u^{p-1} du,$$ where K_3 is constant . . . . . (ii)

Write M_t for the tth moment about the origin of this curve, M_0 being its area.

$$M_0 \cdot M_4 = \int_0^\infty K_3 e^{-u} u^{p+3} du = K_3\left[-e^{-u}u^{p+3}\right]_0^\infty + (p+3)M_3 M_0 \cdot$$
$$= (p+3)M_3 M_0$$
$$= \text{(similarly) } (p+3)(p+2)M_2 M_0 = (p+3)(p+2)(p+1)M_1 M_0 \cdot$$
$$= (p+3)(p+2)(p+1)\,pM_0.$$

$$\therefore M_1 = p, \quad M_2 = p(p+1), \quad M_3 = p(p+1)(p+2),$$
$$M_4 = p(p+1)(p+2)(p+3).$$

Referred to the average $M_2' = p$, $M_3' = 2p$, $M_4' = 3(p^2+2p)$
(p. 251).

$$\kappa_1 = M_3' \div M_2'^{\frac{3}{2}} = 2/\sqrt{p}, \quad \kappa_2 = M_4' \div M_1'^2 = 3\left(1 + \frac{2}{p}\right)$$

Hence the average of $\mu_2 = \dfrac{2m_2}{n}$ (average u) $= \dfrac{2m_2}{n}p =$

$\dfrac{n-1}{n}m_2.$ (Cf. p. 342, note 1); the standard deviation of

$\mu_2 = \dfrac{2m_2}{n}\sqrt{p} = m_2\sqrt{\dfrac{2}{n}\left(1 - \dfrac{1}{n}\right)}.$ (Cf. p. 417 (121), where n is

large), and the measures of skewness and kurtosis are the same for μ_2 and for u, viz.

$$\kappa_1 = \frac{2\sqrt{2}}{\sqrt{(n-1)}}, \quad \kappa_2 = 3\left(1 + \frac{4}{n-1}\right) = 3(1 + \tfrac{1}{2}\kappa_1^2).$$

Notice that equation (i) involves m_2, the second moment of the universe, which is unknown in the ordinary process of sampling.

It is found, however, that the ratio of two values of μ_2 from two samples has a frequency independent of m_2, and this is the basis of " variance " analysis. We have not, however, got over the limitation that the original group is normal, which is very important when n is small. When n is great, however, we know from the analysis pp. 450–2, that the distribution tends to normality, whatever the original group, with standard deviation $\sqrt{\left(\dfrac{1}{n}(m_4 - m_2^2)\right)}$, where m_4 and m_2 are from the original group. In that analysis we should write $a_i = x_i^k - k\mu_{k-1}x_i$, since the moment is computed from the average of the sample.

Also when n is great, Type III tends rapidly to normal

distribution, as is suggested by the values of κ_1 and κ_2 given above. To verify this, write $v = (u - p) \div \sqrt{p}$, thus referring (ii) to its average, with standard deviation as unit.

After some reduction we have $P = K_4 . e^{-v\sqrt{p}} \left(1 + \dfrac{v}{\sqrt{p}} \right)^{p-1} dv,$

where $K_4 = 1/\sqrt{2\pi}$, if p is integral.

$$\log\left\{ e^{-v\sqrt{p}} \left(1 + \frac{v}{\sqrt{p}} \right)^{p-1} \right\} = -v\sqrt{p} + (p-1)\left(\frac{v}{\sqrt{p}} - \frac{v^2}{2p} + \tfrac{1}{3}\frac{v^3}{p\sqrt{p}} - \right)$$

$$= -\tfrac{1}{2}v^2 - \frac{1}{\sqrt{p}}\left(v - \frac{1}{3}v^3 \right),$$

neglecting terms of order $\dfrac{v}{p}$, in comparison with $\dfrac{v}{\sqrt{p}}$.

$\therefore P = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}v^2}\left(1 - \dfrac{\kappa_1}{2}(v - \tfrac{1}{3}v^3) \right)$, the second approxima-

tion to the normal equation (p. 295. Cf. p. 345 (86)), since $\frac{1}{2}\kappa_1 = 1/\sqrt{p}$, and $\kappa_1{}^2$ is negligible.

[See *R.S.S. Journal*, 1931, Irwin, pp. 284–6; *Econometrica*, 1935, Fisher, pp. 353–5; and the references there given, for the development and use of equation (1).]

SUPPLEMENT X.—STANDARD DEVIATIONS OF PERCENTILES, MEAN DEVIATION, AND MEAN DIFFERENCE.

(With p. 417.)

Percentiles.

Let $y = f(x)$ be a continuous frequency curve ranging from h to k, such that $\int_h^k f(x)\, dx = 1$.

Write $p = \int_h^x f(x)dx$ = proportion of cases below value x, so that x is the $(100p)$th percentile of the distribution.

Then $dp = f(x)dx = ydx$, and the increase in the value of x corresponding to a small increase in p is approximately

$$\Delta x = \frac{1}{y}\Delta p \qquad . \quad . \quad . \quad . \quad . \quad (1)$$

In a sample of n objects, let the observed proportion below x be $p + \Delta p$. In repeated samples the average of Δp is zero, and mean

$$(\Delta p)^2 = p(1 - p)/n \quad . \quad . \quad . \quad . \quad . \quad (2)$$

Then, if n is great, the relation of the error in x obtained[*] from the sample to Δp is given by (1), and therefore from (2) $_p\sigma_x{}^2 = p(1 - p)/ny^2$, where $_p\sigma_x$ is the standard deviation of the errors of x deduced from the sample.

Thus for the $(100p)$th percentile $_p\sigma_x = \frac{1}{y}\sqrt{\left(\frac{p(1 - p)}{n}\right)}$ is the standard deviation in the scale reading.

If the frequency curve is normal $y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$.

In normal curve

Percentile. p.	$\sqrt{(p(1 - p)}$	$\dfrac{x^*}{\sigma}$	$y\sigma$	$_p\sigma_x \div \dfrac{\sigma}{\sqrt{n}}$ †
·1 or ·9	·3	1·281	·176	1·71
·2 or ·8	·4	0·842	·280	1·43
Quartiles	·433	0·6745	·317	1·36
·3 or ·7	·458	0·5233	·347	1·32
·4 or ·6	·490	0·2533	·386	1·27
Median	·5	0	·399	1·25

Thus the standard deviation of the median is $1·25\sigma/\sqrt{n}$ approx., that is very nearly $\frac{5}{4}$ of the standard deviation of the average (p. 289 (38)).

To find the standard deviation of the interval observed between percentiles $100p$ and $100q$ symmetrically placed, arrange the divisions thus

Proportions found $p + dp_1$ | $q - p - dp_1 - dp_2$ | $p + dp_2$
Range h to x_1 | x_1 to x_2 | x_2 to k.

Then mean $dp_1 . dp_2 = - p^2/n$, from p. 419, line 8.

The interpercentile distance is $x_2 - x_1 = u$ (say), and we can argue as before that $y . du = dp_1 + dp_2$, where y is the ordinate corresponding to either percentile.

$$\therefore y^2\sigma_u{}^2 = \text{mean}\{(dp_1)^2 + (dp_2)^2 + 2dp_1 . dp_2\} = (pq + pq - 2p^2)/n$$

$$\therefore \sigma_u = \frac{1}{y}\sqrt{\left\{\frac{2}{n}(p(1 - 2p))\right\}},$$

where σ_u is the standard deviation required.

* Value of z corresponding to $\frac{1}{2} - p$ on p. 271.

† $\sqrt{(p(1 - p))} \div y\sigma$. All values in the last three colums are approximate.

For example, for the interquantile range $\sigma_u = 1/2y\sqrt{n}$.

For the probable error (ρ), that is half the interquantile range, $\sigma_p = 1/4y\sqrt{n}$, $= \cdot786\sigma/\sqrt{n}$, if the distribution is normal, $= 1\cdot18\rho/\sqrt{n}$, since $\rho = \cdot6745\sigma$ (p. 272 (26)).

while, also in the normal curve, the standard deviation of the standard deviation is given by

$$\sigma_\sigma = \sigma/\sqrt{2n} = \cdot707\sigma/\sqrt{n} \text{ (p. 420 (127)).}$$

[Compare Yule—*An Introduction to the Theory of Statistics*, pp. 337–8 and 343.]

Mean Deviation.

Consider the mean deviation, η_u, about a central position u.

The deviations of a frequency group, all taken positively, form a frequency group ranging upwards from zero, whose average is η_u, and second moment about zero is given by

$\frac{1}{N} S(x - u)^2 = \sigma^2 + (u - \bar{x})^2$, where σ and $\bar{x}$ are the standard deviation and average of the original group.

∴ Standard deviation of the new group is

$$\sqrt{\{\sigma^2 + (u - \bar{x})^2 - \eta_u^2\}} = s,$$

say. Hence, in repeated samples of n each, the standard deviation of the average, that is of η_u, is $s/\sqrt{n}$ (see p. 289 (38)).

If the deviations are taken originally from the average, $u = \bar{x}$, and $_a\sigma_\eta = \sqrt{\{(\sigma^2 - \eta^2)/n\}}$.

For the normal curve $\eta = \sigma\sqrt{(2/\pi)}$, (p. 269 (24)), and $_a\sigma_\eta = \cdot603\sigma/\sqrt{n} = \cdot756\eta/\sqrt{n}$, approximately.

In the second approximation to the normal curve the distance from average to median is of order $1/\sqrt{n}$ (p. 444 (145)), and when $1/n$ is neglected, the standard deviation of the mean deviation from the median is very nearly the same as $_a\sigma_\eta$.

Mean Difference (see pp 114–5).

The standard deviation of the mean difference for the normal curve is given by $\sigma_g = \dfrac{\sigma}{\sqrt{n}} \left(\dfrac{4}{3} - \dfrac{8(2 - \sqrt{3})}{\pi} \right)$

$= \cdot807\sigma/\sqrt{n} = \cdot715g/\sqrt{n}$ approx., since $g = \eta\sqrt{2} = 2\sigma/\sqrt{\pi}$.

[Bowley and Wold Congrès International des Mathématiciens, Oslo, 1936].

SUPPLEMENT XI.—STANDARD DEVIATION OF THE CORRELATION
COEFFICIENT.

(With p. 422.)

It is worth while to evaluate in a simple case the complete
formula for σ_r given in the last line of p. 422.

Let $x = u_1 + u_2 + \ldots + u_m$, $y = v_1 + v_2 + \ldots + v_m$, all
the variables being measured from the averages.

Let each u and v have the same standard deviation σ and
fourth moment μ_4, and with $\mu_4 = (3 + \epsilon)\sigma^4$, where ϵ is zero if
the curve is normal.

Let the u's be uncorrelated and the v's be uncorrelated.

Let u_s and v_t be uncorrelated, so that

Mean $u_s v_t = 0 = $ mean $u_s v_t^3$.   Mean $u_s^2 v_t^2 = \sigma^4$ $s \neq t$

Let each pair (u_t, v_t) be correlated by the same frequency
distribution, so that mean $u_t^3 v_t = (3\rho + d)\sigma^4 = $ mean $u_t v_t^3$
and mean $u_t^2 v_t^2 = (2\rho^2 + 1 + \delta)\sigma^4$.
where mean $u_t v_t = \rho\sigma^2$.

d and δ are zero if the correlation surface is normal.

(p. 362 (106)).

Then with the notation of p. 422
$$\lambda = M_{20}* = m\sigma^2 = M_{02} = \mu,$$
where m is the number of independent u's or v's.

$M = M_{11}* = $ sum mean $u_t v_t = m\rho\sigma^2 = \rho\lambda$, and $r = \rho$

$\lambda_4 - 3\lambda^2 = M_{40}* - 3\lambda^2 = m(\mu_4 - 3\sigma^4)$, from p. 292

$$= m\sigma^4\epsilon = \frac{1}{m}\lambda^2\epsilon = \mu_4 - 3\mu_2^2$$

$M_{31}* = $ mean $(Su_t^3 \times Sv_t)$
$= S$ (mean $(u_t^3 v_t)$) $+ 3 S$ (mean $u_t^2 . u_t v_t$)
$+$ zero terms.
$= m(3\rho + d)\sigma^4 + 3 m(m - 1)\sigma^2 . \rho\sigma^2$

$\therefore \dfrac{M_{31}}{\lambda M} = 3 + \dfrac{1}{m} . \dfrac{d}{r} = \dfrac{M_{13}}{\mu M}.$

$M_{22}* = $ mean $(Su_t^2 \times Sv_t^2)$
$= S$ (mean $(u_t^2 v_t^2)$) $+ 2 S$ (mean $(u_s^2 v_t^2)$
$+ 4 S$ (mean $(u_t u_s v_s v_t)$) $+$ zero terms.
$= m(2\rho^2 + 1 + \delta)\sigma^4 + m(m - 1)\sigma^4 + 2m(m - 1)\rho^2\sigma^4$

$\therefore \dfrac{M_{22}}{\lambda^2} = 2r^2 + 1 + \dfrac{\delta}{m}.$

* These are obtained by multiplying appropriate powers of x and y and
taking the means

Substitute these values in the formula (p. 422)

$$\sigma_r{}^2 = \frac{r^2}{n}\left\{\left(2 + \frac{1}{r^2} + \frac{3}{4} + \frac{3}{4} - 3 - 3 + r^2 + \frac{1}{2}\right)\right.$$
$$\left. + \frac{1}{m}\left(\frac{\delta}{r^2} + \frac{\epsilon}{4} + \frac{\epsilon}{4} - \frac{d}{r} - \frac{d}{r} + \frac{\delta}{2}\right)\right\}$$
$$= \frac{1}{n}\left\{(1 - r^2)^2 + \frac{1}{m}\left(\delta(1 + \frac{r^2}{2}) + \epsilon\frac{r^2}{2} - 2dr\right)\right\}$$

Hence, if m (the number of independent elements) is considerable, or if the frequency curves and surfaces of the u's, v's and pairs of $u_i v_i$ approach normality, the correction to the formula $(1 - r^2)/\sqrt{n}$ is slight, and in any case is of the order $1/2m$. But, if m is small, the correction may be considerable and be either positive or negative.

Other cases, of a slightly more complex kind, are worked out in a Note in the *Journal of the American Statistical Association*, 1928, pp. 31–4, of which the above is a modified version.

SUPPLEMENT XII.—THE METHOD OF CONFIDENCE BELTS.

(With pages 412 *seq.*)

Many statisticians wish to be independent of any hypothesis about *à priori* probability, when they draw inferences from the results of sampling. The method of confidence belts or fiduciary limits, introduced by Prof. R. A. Fisher, has this purpose, and we proceed to describe one aspect of it in the simplest case.

Write $\Pi(p,x) = {}_nC_x \cdot p^x q^{n-x}$ for the chance of obtaining x white balls in n independent selections from an urn containing white and black in the proportion $p : q(p + q = 1)$, where p remains unaffected by the drawing. Let pn be sufficiently great to allow us to write $\Pi = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$, where $z = (x - pn)/\sqrt{pqn}$.

Choose some limit, say ·05, and find from the normal Table (p. 271) the value of z that makes $2\int_z^\infty \Pi \cdot dz = ·05$, approx. This value is 1·96 ..., so that $\frac{x}{n} = p \pm 1·96\sqrt{\frac{pq}{n}}$.

Ih·any experiment n is known. Take the case $n = 100$, and compute x/n for a series of values of p.

p.	x/n.	p.	x/n
0	0	1·0	1
·1	·1 $\pm$ ·06	·9	·9 $\pm$ ·06
·2	·2 $\pm$ ·08	·8	·8 $\pm$ ·08
·3	·3 $\pm$ ·09	·7	·7 $\pm$ ·09
·4	·4 $\pm$ ·096	·6	·6 $\pm$ ·096
·5	·5 $\pm$ ·1		

Take rectangular axes, OX, OP and plot the values of p and x/n. We obtain two curves (in this case ellipses) enclosing a space, which is called a " confidence belt."

Suppose x to be found be 20 in one experiment. Mark OM $=$ ·2, and through M draw a perpendicular to OX to intersect the ellipses in L and K.

ML and MK are the values of p given by

$$·2 = p \pm ·196 \sqrt{pq/n},$$

that is, ·13 and ·29 approx.

Now suppose that there are a number of urns from which the drawing may have been made, and that the proportion of urns, in which the ratio of white to black is $p_1 : q_1$, is P_1, so that $S(P_1) = 1$, the sum being extended over all values of p from 0 to 1. We know nothing about P_1; it may vary continuously in any way, or be located at one or more particular points.

The chance of choosing an urn ($p_1 : q_1$) and drawing x white balls is $P_1 \times \Pi(p_1 x) = Z$. The sum of Z over all values $p = 0$ to 1, and $x = 0$ to n is unity.

Conceive a surface determined by the extremities of lines perpendicular to the plane XOP equal to Z, and suppose a vertical cylinder through the ellipses bounding the belt of confidence.

This cylinder includes ·95 of the area of every vertical section perpendicular to OP. For example, it includes the values of Z standing on HJ, the plane through AB, drawn through $p_1 = $ OA $= $ ·4 (AH $= $ ·304, AJ $= $ ·496, from the little table above).

The area of this section is $P_1 \times S_{x=0}^{n} \Pi(p_1 x) = P_1 \times 1 = F_1$,

which may have any value from o to 1, but in every case the proportion of that value included is ·95. Therefore the proportion of the whole volume bounded by the surface which is included in the cylinder is ·95, however P_1 may be distributed.
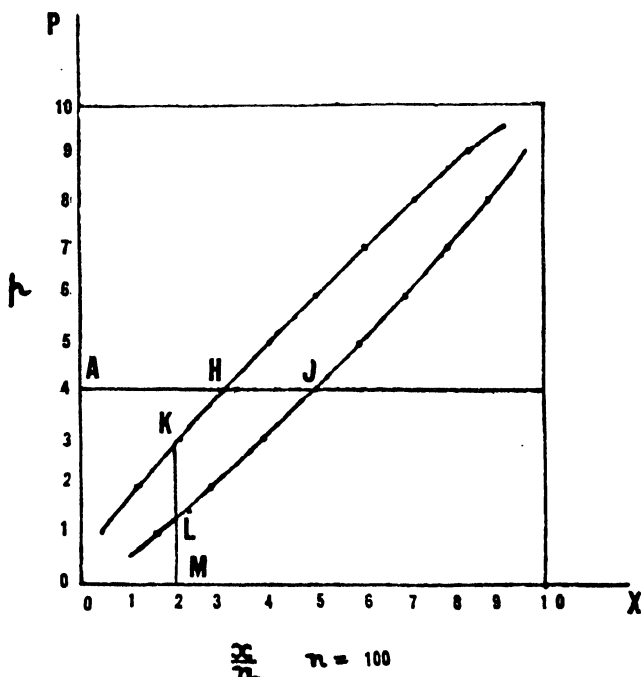
DIAGRAM E.

For any one drawing of n balls in which $x \left(\dfrac{x}{n} = OM \right)$ white balls are found, obtain the points L and K.* From this result write down the hypothesis, "p the unknown proportion in the urn from which the drawing took place is between ML and MK." We have no means of testing the accuracy, or probability of the truth, of this hypothesis in any one case. But if drawings are repeated an indefinitely large number of times, till we may assume that all the urns are engaged in the proportions P_1's, and from each of them the frequency of the

* $KL = l\sqrt{\{4n \cdot \dfrac{x}{n}\left(1 - \dfrac{x}{n}\right) + l^2\} \div (n + l^2)}$, where l is written for the 1·96 in the equation above. KL diminishes as $\sqrt{n}$ increases. Its maximum is at $x = \frac{1}{2}n$, for any given value of n and equals $(n/l^2 + 1)^{-\frac{1}{2}}$.

resulting x's is given by the appropriate Π, then the theoretical distribution of $Z = P_1 \times \Pi$ will in the very long run tend to be reached. Since 95 per cent. of all the results are within, the confidence belt, the hypothesis written above will be justified in 95 per cent. of an indefinitely large number of experiments, whatever the distribution of the urns.

But we can make no statement whatever about the probability of p from a single drawing or trial, however great n may be, if we have no reasonable knowledge of, or reasonable assumption about, the universe of urns. (Of course a series of confidence belts can be drawn corresponding to various fiduciary limits; ·05 is taken only for numerical illustration).

A result of this kind has its use when a great number of experiments have been made, and we do not depend on the accuracy of the individual estimates. But if we have only one sample, such as was obtained in the *New Survey of London Life* and Labour, any action (such as the supply of sufficient milk to children) must be based on that one; any confidence we have in our results is based on n being large, and on the validity of the assumption as to *à priori* probability indicated on p. 414. The main assumption there is that the frequency of the occurrence of " urns," with proportions differing significantly from the central region, is not overwhelmingly great, and this may in many cases be known from general experience of the " populations " with which we are concerned.

[The method of approach in the above and the framework of the diagram are largely based on " The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial," *Biometrika*, Dec. 1934, C. J. Clopper and E. S. Pearson. The reader is also referred to Dr. Neyman's paper in the *R.S.S. Journal*, 1934, pp. 589–93. The general ideas came from Professor R. A. Fisher's work, but it is not to be assumed that he would accept the exegesis here offered without at least serious qualifications.]

SUPPLEMENT XIII.—THE TEST OF "GOODNESS OF FIT."

(With page 431.)

Since it is now difficult to refer to Pearson's analysis given in 1900, it may be well to indicate the steps in it without attempting a complete proof.

From the expression for X^2, viz. : $\dfrac{e_1^2}{m_1} + \dfrac{e_2^2}{m_2} + \ldots + \dfrac{e_n^2}{m_n}$, one of the e's may be eliminated, since $e_1 + e_2 + \ldots + e_n = 0$. X^2 is then a homogeneous quadratic expression in $\overline{n-1}$ quantities, and can therefore be expressed by linear transformations as the sum of $\overline{n-1}$ squares, and written

$$X^2 = u_1^2 + u_2^2 + \ldots + u_{n-1}^2$$

Then $P = I_X \div I_0$, where

$$I_X = \iint \ldots K e^{-\frac{1}{2}(u_1^2 + u_2^2 + \ldots u_{n-1}^2)} \, du_1 du_2 \ldots du_{n-1}$$

integrated over the range of u's which make $S(u^2)$ as great as an assigned X^2.

The process of integration is similar to that for finding the volume of a sphere.

Take $n = 4$.

Write $u_1 = X \sin\theta \cos\phi$, $u_2 = X \sin\theta \sin\phi$, $u_3 = X \cos\theta$.

Then $u_1^2 + u_2^2 + u_3^2 = X^2$.

$$I_X = \int_X^\infty \int_{0=-\frac{\pi}{2}}^{\frac{\pi}{2}} \int_{\varphi=0}^{2\pi} K e^{-\frac{1}{2}X^2} . d\phi . \sin\theta . d\theta \; X^2 dX$$

$$= 4\pi K \int_X^\infty e^{-\frac{1}{2}x^2} X^2 . dX.$$

By analogy in $\overline{n-1}$ dimensions, $I_X = K_1 \int_X^\infty e^{-\frac{1}{2}x^2} X^{n-2} . dX.$

Integrate this by parts

$$\frac{1}{K_1} . I_X = \left[-e^{-\frac{1}{2}x^2} . X^{n-3} \right]_X^\infty + (n-3)\int_x^\infty e^{-\frac{1}{2}x^2} X^{n-4} dX$$

$$= e^{-\frac{1}{2}x^2} . X^{n-3} +$$

$(n-3)\left\{ e^{-\frac{1}{2}x^2} X^{n-5} + (n-5)\int_X^\infty e^{-\frac{1}{2}x^2} X^{n-6} dX \right\}$, and so on.

If n is even, the final integral is $\int_X^\infty e^{-\frac{1}{2}x^2} dX = f(X)$, say.

Then $\dfrac{1}{K_1} \cdot I_X = e^{-\frac{1}{2}X^2}\Big\{ X^{n-3} + (n-3)X^{n-5}$

$\qquad\qquad + (n-3)(n-5)\, X^{n-7} + \ldots + (n-3) \ldots 5.3.X\Big\}$

$\qquad\qquad + (n-3)\ \ldots\ 3 \cdot 1 f(X)$

and $\dfrac{1}{K_1} \cdot I_0 = (n-3) \ldots 3 \cdot 1\sqrt{\dfrac{\pi}{2}}$, since $f(0) = \sqrt{\dfrac{\pi}{2}}$.

Reverse the order of the expansion of I_X, and divide by I_0, and we obtain the equation at the bottom of p. 430.

When n is odd the final steps are simpler and we have the equation (131).

Write $\frac{1}{2}X^2 = v$ and $n = 2m + 1$.

The formula for n odd becomes

$$P = e^{-v}\Big(1 + v + \dfrac{v^2}{2!} + \ldots + \dfrac{v^{m-1}}{(m-1)!}\Big)$$

$$= P_0 + P_1 + \ldots \ P_{m-1}, \text{ in the notation of p. 285 (29).}$$

In such a series continued indefinitely, the greatest term is $e^{-v} \cdot v^v/(v)!$, and since the series is the limit of a binomial expansion, the mode (given by this term) may be expected to be less than the median by approximately $\frac{1}{3}\kappa\sigma$, (p. 444 (145)) which $= 1$, when $\sigma = \sqrt{v}$, $\kappa = 1 \div \sqrt{v}$ (p. 285 (30) (31)).

If $v = m - 1$, we have the greatest term, and $v = m$ gives the median approximately.

Hence $P = \frac{1}{2}$ for some value of $\frac{1}{2}X^2$ between $m - 1$ and m, that is for some value of X^2 between $n - 3$ and $n - 1$.

This is a rough explanation of the cause of the relation stated in the first line below the Table on p. 431. It can, of course, be verified by a full Table of n, X^2 and P.

Effect of Restrictions in Sampling.

The linear relations between the e's (p. 429 (130)) has the same effect as reducing the number of independent variables from n to $\overline{n-1}$, and results in the index $\overline{n-2}$ in the expression for P. Every additional linear relation reduces the index one unit further, since it makes possible the elimination of one more e. If there are c additional relations the n at the head of the first column of the Table on p. 431 must be interpreted as $\overline{n-c}$; e.g. if $n = 10$ and $c = 2$, $P = \cdot54$ corresponds to $X^2 = 6$.

Thus in a contingency table (p. 373) it is often·assumed· that the sub-totals, n_1, n_2 . . . ., m_1, m_2 . . . are not subject to variation. When this is so, and l and c are the numbers of lines and columns in the table, we have $\overline{c-1}$ and $\overline{l-1}$ additional linear relations between certain of the e's. If the sub-totals are variable, n, the number of compartments, $= c \times l$, and the index of X^2 is $cl - 2$; with the restrictions it is $cl - 2 - (c-1) - (l-1) = cl - c - l$. (*R.S.S. Journal*, 1922, R. A. Fisher, p. 88).

Thus in the case $c = 2$, $l = 2$, $P = \int_{X}^{\infty} \sqrt{\frac{2}{\pi}} . e^{-\frac{1}{2}X^2} . X^0 dX$,

where $X^2 = \dfrac{N^3 x^2}{n_1 n_2 m_1 m_2}$, as in fact is given on p. 372.* If, how-

ever, the ratios of the sub-totals were not known, but were based on the observations of one random group taken from a larger universe we should have X^2 instead of X^0.

Here we are in the difficulty that unless we know the proportions in the universe, we have not the data for calculating α, β . . . on which X^2 depends.

A similar question arises when we calculate the constants of a frequency curve from the sample to which it is to be fitted. Thus in the example on pp. 274–5 we may assume $p = \frac{1}{2}$, $\sigma = 3\cdot535$, and the only doubt about the test is how fine the grading of the results should be; but in the examples on pp. 304–6, we can only determine $\bar{x}$ and σ from the observations.

A considerable difference of opinion has existed on the question whether this process involves additional restrictions. It appears to depend on the hypothesis made as to the method of repeated sampling, where in fact we have only one sample.

The different hypotheses are :—

1. Suppose a great number of samples to be taken from the same universe, and that they are not restricted to have the average and other moments of the given sample. Then there is no restriction, but we do not know the universe. We can compute X^2 and P for any universe we like to define; among the results from a universe of the given form (such as normality) which is to be tested, we may choose that which minimises

* But there the chance of a positive deviation (or excess) is taken, so that we have $\frac{1}{2}$P.

X^2; then we could say that the chance that the sample would arise from the universe which fitted it best is P, with no restriction. In Professor Karl Pearson's words "what we actually do is to replace the accurate value of X^2, which is unknown to us and cannot be found, by an approximate value," when we select the constants on this or any other principle. If the number of objects in the sample is large X^2 will vary, under ordinary choice, only by a quantity comparable with n^{-1}, which in fact has already been neglected (p. 454).

2. It is supposed that samples are drawn again and again and X^2 is computed for each sample by equating the moments in the universe to those in that sample. Then the samples are not supposed to be drawn from an invariable universe. In such a case the analysis of pp. 429–31 does not apply, for the m's are variable; we have a number of single examples for a series of values of X^2.

3. The samples are drawn from the same universe, and X^2 is calculated subject to one or more linear restrictions. The index in the integral is then reduced. The resulting P = function X shows the distribution of chances of samples restricted to definite moments or other constants, that is from a universe in which certain quantities are taken as known.

Hypothesis 3 is certainly appropriate in a contingency table where the sub-totals are in fact known. In other classes of cases we are entitled, in my opinion, to choose the universe in any way we please, with or without reference to values based on the sample, and without modifying the universe. The reader may perhaps be helped to form a judgment by considering Gibrat's graphic method (p. 471). If we choose x_0 so as to get optically the least curvature in the line representing the observations and then read off a and b and compute X^2, have we any reason to say that the sample is taken under the condition that x_0, a and b restrict it? Or is that an unfair way of stating hypothesis 3 and its results?

[See Fisher, *loc. cit.*, and in *Economica*, 1923, pp. 139–147; Pearson, *Biometrika*, 1922, pp. 186–91; Bowley and Connor, *Economica*, 1923, pp. 1–9, and the references there given.]

INDEX

ELEMENTS OF STATISTICS

PERSONAL INDEX